

# **Lost in Transition**

## **- Genetic, Transcriptomic and Breeding Aspects of Metabolic Robustness in Dairy Cows**

Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Fakultät für Agrarwissenschaften

der Georg-August Universität Göttingen

vorgelegt von

Ngoc-Thuy Ha

geboren am 12.01.1988 in Ben-Tre (Vietnam)

Göttingen, Juni 2016

1. Gutachter: Prof. Dr. Henner Simianer

2. Gutachter: Prof. Dr. Martin Schlather

Tag der mündlichen Prüfung: 23.06.2016

## TABLE OF CONTENTS

Summary.....	4
Zusammenfassung.....	7
Chapter 1 General Introduction.....	11
The Transition Period of Dairy Cows .....	12
Genome-wide Association Studies.....	13
RNA Sequencing and Differential Gene Expression Analysis .....	16
Reaction Norm Models.....	19
References.....	21
Chapter 2 Gene-based Mapping and Pathway Analysis of Metabolic Traits in Dairy Cows .....	27
Chapter 3 Whole Liver Transcriptome Analysis Reveals Novel Factors Involved in the Metabolic Adaptation of the Transition Cow .....	58
Chapter 4 Characterizing Metabolic Robustness of Dairy Cows During Early Lactation Using Reaction-Norm Models.....	91
Chapter 5 General Discussion .....	112
Overview.....	113
Genome Annotations – the Bridge Between Genome and Functional Biology?.....	114
RNA-seq – the Curse of Small Sample Sizes .....	115
Reaction Norm Model – the Model Assumptions.....	117
Outlook .....	118
References.....	119

## SUMMARY

The transition phase of dairy cows is marked by severe metabolic stress resulting from a discrepancy of a high energy demand for rapidly increasing milk production and limited feed intake. Here, a failure in metabolic adaptation results in an increased susceptibility to health problems. However, even under the same environmental factors and production level, the variability of how each cow deals with metabolic load is substantial, leading to the hypothesis that there might be an underlying genetic basis. The main goal of this thesis is to study this genetic basis from different points of view.

In **chapter 2**, we studied the genetic basis of the metabolic adaptation by means of auxiliary phenotypes best characterizing the adaptive process. Blood samples were taken from 178 cows at three critical stages: T1 = week 3 *ante-partum* (no metabolic load); T2 = week 4 *post-partum* (lactating and high metabolic load), and T3 = week 13 *post-partum* (lactating and low metabolic load). Plasma concentrations of non-esterified fatty acids (NEFA), beta-hydroxybutyrate (BHBA) and glucose – metabolites characterizing the metabolic status and adaptability – were measured at T1, T2, and T3. All cows were genotyped with the Illumina HD Bovine BeadChip. After quality control, the remaining 601,455 SNPs were annotated to genes (Ensembl) and pathways (KEGG). For each gene and phenotype, we performed a modified score test based on a linear regression model with all SNPs in the gene as explanatory variables, while taking into account possible environmental or breed effects. The results were used to identify pathways enriched for significant genes using a weighted Kolmogorov-Smirnov test. We found 99 genes significantly associated with the three metabolites. For each metabolite, we found genes that are significant at T2 but not at T1 and T3 or vice versa. This strongly suggests those genes to be involved in the adaptive regulation. We further identified three pathways ('steroid hormone biosynthesis', 'ether lipid metabolism' and 'glycerophospholipid metabolism') jointly affecting the metabolites.

Even though NEFA, BHBA and glucose are important metabolites describing the metabolic adaptation, they may not be fully indicative for the whole process. In order to obtain a more complete picture of the metabolic adaptation, **in chapter 3** we conducted a whole transcriptome analysis of the liver, since it is the key organ controlling and regulating the metabolic adaptation. Liver samples were taken from 6 cows at 3 time-points: T1 = week 3 *ante-partum*; T2 = week 2 *post-partum*; T3 = week 3 *post-partum*. Using RNA-seq, we studied the transcriptomic profile of the transition cow before and after lactation. We performed a differential gene expression analysis (DGE) and a combination of the gene-set enrichment analysis and perturbation analysis for pathways (KEGG database). Among the ~10,000 expressed genes, we discovered ~1,000 genes to be significantly differentially expressed (FDR < 5%), of which ~43% and ~16% are linked to lipids and oxidative stress, respectively, but only ~6% to the glucose metabolism (GO database). The combined pathway analysis further revealed seven pathways to be significantly associated with the hepatic changes of the transition cow, including ‘adipocytokine signaling pathway’ and ‘steroid hormone biosynthesis’. The DGE and pathway analysis demonstrated that major hepatic changes from late pregnancy to early lactation relate to gluconeogenesis and fat mobilization (‘adipocytokine signaling pathway’). We further found indications for immunological changes (*GPX3*, ‘steroid hormone biosynthesis’ and the associated *CYP* and *UGT* transcripts) that may contribute to the impaired immune system of dairy cows during the transition period. The outcome of this study provides new insights into the metabolic adaptation which should be more closely investigated in future studies.

The main results of **chapter 2** and **chapter 3** indicate the possible existence of a genetic component of the metabolic adaptation. In other words, whether a cow is able to adapt successfully or not may be partly determined by her genetic set up. Therefore, in **chapter 4**, we hypothesize that some cows are genetically less well suited to cope with this metabolic stress than others, leading to adverse follow-up effects on longevity. We thus designed a

reaction norm sire model linking the functional lifetime to the metabolic challenge in early lactation. To this end, we used either the sum of the milk yield or the accumulated fat/protein ratio of the first three test days to define a measure for metabolic load that a cow has to face during her early lactation. To assess the genetic merit and heritability of the metabolic adaptability, we defined a pedigree-based random regression sire model, in which a random regression term was estimated for each sire to reflect the genetic component of the reaction to the challenge. The model was fitted to data of ~580,000 daughters of ~5,000 Brown Swiss bulls with at least 10 daughters with records. We found the sire variance for the slope of the random regression to be significantly different from zero for both challenge variables, suggesting a genetic component for the ability to cope with metabolic stress. The results of the study show that the ability to cope with metabolic stress in the transition phase clearly has a genetic component and could be used to breed metabolically robust dairy cows.

In conclusion, by assessing the genetic basis of the metabolic adaptation from a genetic, transcriptomic as well as a breeding point of view, we found strong evidence supporting this hypothesis. Despite the high complexity of metabolic adaption, we found several genetic factors affecting the adaptability even across different studies. The identified factors as well as the newly developed measure for the metabolic robustness are not only a valuable contribution to the understanding of the transition cow, but an effective tool for the dairy industry to breed for metabolically robust dairy cows.

## ZUSAMMENFASSUNG

Während der Transitphase unterliegt die Milchkuh einer enormen metabolischen Belastung. Durch den ansteigenden Energiebedarf für die einsetzende Milchproduktion und die begrenzte Futteraufnahmekapazität entsteht ein kritisches Energiedefizit. Hierbei führt die unzureichende metabolische Anpassungsfähigkeit zu einem erhöhten Risiko des Auftretens einer Stoffwechselstörung. Unabhängig von Umweltbedingung oder Produktionsniveau zeigt sich, dass einzelne Kühe sehr unterschiedlich mit dieser Belastung umgehen können. Dies hat zu der Hypothese geführt, dass die Anpassungsfähigkeit des Stoffwechsels auch einen genetischen Hintergrund hat. Hauptziel dieser Arbeit ist es, die genetische Basis zu untersuchen.

In **Kapitel 2** wurde die genetische Grundlage der metabolischen Anpassung mittels Hilfsmerkmalen untersucht, von denen wir annehmen, dass sie den adaptiven Prozess am besten charakterisieren. Es wurden Blutproben von 232 Kühen zu drei Zeitpunkten der Transitperiode entnommen: T1 = 3 Wochen antepartum (keine metabolische Belastung), T2 = 4 Wochen postpartum (laktierend und hohe metabolische Belastung) und T3 = 13 Wochen postpartum (laktierend und geringe metabolische Belastung). Um den Stoffwechselstatus und die Anpassungsfähigkeit zu charakterisieren, wurden Plasmakonzentrationen von nicht-veresterten Fettsäuren (NEFA), beta-Hydroxybutyrat (BHBA) und Glukose, drei wichtigen Metaboliten des Energiestoffwechsels, an jedem der genannten Zeitpunkte bestimmt. Alle Kühe wurden mit dem Illumina HD Bovine BeadChip genotypisiert. Nach der Qualitätskontrolle verblieben 601'455 SNPs für die Analyse. Für diese wurden Gen (Ensembl) und Pathway Informationen (KEGG) annotiert. Basierend auf einem linearen Regressionsmodell, das alle SNPs eines Genes als erklärende Variablen beinhaltet, und unter Berücksichtigung möglicher Umwelteffekte und Populationsstrukturen, führten wir für jedes Gen und jeden Phänotyp einen modifizierten Score-Test durch. Mit einem nachfolgenden gewichte-

ten Kolmogorov-Smirnov-Test wurden Pathways identifiziert, die mit signifikanten Genen angereichert waren. Insgesamt fanden wir 99 signifikante Gene, die mit mindestens einem der drei Metaboliten assoziiert waren. Für jeden Metaboliten existierten Gene, die nur bei T2 aber nicht bei T1 und T3 signifikant waren oder umgekehrt, was darauf hindeutete, dass diese Gene einen Einfluss auf die metabolische Anpassungsfähigkeit haben könnten. Des Weiteren wurden drei Pathways (‘steroid hormone biosynthesis’, ‘ether lipid metabolism’ und ‘glycerophospholipid metabolism’) identifiziert, die jeweils einen Effekt auf alle drei Metaboliten haben.

Obwohl NEFA, BHBA und Glukose wichtige Metaboliten für die metabolische Adaptation darstellen, bilden sie den gesamten Prozess der Adaptation nicht umfassend ab. Um ein vollständigeres Bild der metabolischen Anpassung zu erhalten, führten wir eine Transkriptom-Analyse der Leber durch, einem Schlüsselorgan der Umsetzung metabolischer Prozesse, die in Kapitel 3 dargestellt wird. Leberproben wurden von 6 Kühen zu 3 verschiedenen Zeitpunkten entnommen: T1 = 3 Wochen antepartum; T2 = 2 Wochen postpartum; T3 = 3 Wochen postpartum. Mit Hilfe der RNA-sequenzierungstechnologie untersuchten wir das transkriptomische Profil der 6 Kühe vor und nach der Abkalbung. Wir führten eine differentielle Genexpressionsanalyse (DGE) und eine Kombination aus der Gene-Set-Enrichment-Analyse und Perturbationsanalyse für Pathways (KEGG Datenbank) durch. Unter den ~10.000 exprimierten Genen waren etwa 1.000 Gene signifikant differentiell exprimiert (FDR <5%). Von diesen konnten 43% mit Lipiden, 16% mit oxidativem Stress, aber nur 6% mit dem Glukosestoffwechsel in Verbindung gebracht werden (GO Datenbank). Die kombinierte Pathway-Analyse ergab sieben signifikante Pathways, die mit der hepatischen Adaptation der Transitkuh assoziiert waren, einschließlich der Pathways ‘adipocytokine signaling pathway’ und ‘steroid hormone biosynthesis’. Die DGE und Pathway-Analyse zeigten, dass die wesentlichen hepatischen Veränderungen in der Transitphase sich auf die Gluconeogenese und Fettmobilisierung beziehen. Des Weiteren fanden



wir Indizien für immunologische Veränderungen (*GPX3*, ‚steroid hormone biosynthesis‘ und die damit verbundene *CYP* und *UGT*-Transkripte), die zum geschwächten Immunsystem von Milchkühen während der Übergangszeit beitragen könnten. Das Ergebnis dieser Studie liefert neue Einblicke in den Komplex der metabolischen Adaptation, der in zukünftigen Studien weiter untersucht werden sollte.

Die Ergebnisse der **Kapitel 2** und **3** deuten auf die Existenz einer genetischen Basis der metabolischen Adaptation hin. Ob der Stoffwechsel einer Kuh erfolgreich in der Lage ist, sich an die Belastungen der Transitphase anzupassen oder nicht, kann zum Teil aus ihrer genetischen Veranlagung abgeleitet werden. Aus diesem Grund überprüften wir in **Kapitel 4**, ob es eine genetisch bedingte Variabilität gibt, die erklärt, wie Kühe mit der metabolischen Stresssituation in der Frühlaktation umgehen. Dabei stellen wir die Hypothese auf, dass der metabolische Stress bei weniger robusten Kühen negative Auswirkungen auf die funktionale Nutzungsdauer hat, während dies bei robusteren Kühen ausbleibt. Zu diesem Zweck haben wir ein Reaktionsnorm-Vatermodell aufgestellt, in dem wir die funktionelle Lebensdauer mit der metabolischen Belastung in der Frühlaktation verknüpfen. Als Indikatoren für die metabolische Belastung haben wir die kumulierte Milchleistung bzw. das kumulierte Fett/Protein-Verhältnis der ersten drei Herdentesttage genommen. In dem Reaktionsnorm-Vatermodell wurde für jeden Bullen eine zufällige Regressionsgerade bestimmt, deren Steigung beschreibt, wie sensitiv oder robust die Töchter dieses Bullen auf die metabolische Belastung reagieren und somit ein neues Maß für die metabolische Robustheit darstellt. Das Modell wurde an die Daten von ~580.000 Töchtern von ~5.000 Bullen der Rasse Brown Swiss angepasst. In beiden Modellen, in denen jeweils die kumulierte Milchleistung oder das kumulierte Fett/Protein-Verhältnis als Challenge-Variable benutzt wurden, konnte die Verwandtschaftsstruktur der Väter einen signifikanten Teil der Varianz der Regressionskoeffizienten erklären, damit sehen wir unsere Hypothese von der genetischen Komponente der metabolischen Adaptationsfähigkeit bestätigt.

Abschließend lässt sich sagen, dass wir durch die Betrachtung der verschiedenen Aspekte der metabolischen Adaptation (genetische, transkriptomische und züchterische) starke Hinweise auf einen genetischen Hintergrund gefunden haben. Trotz der hohen Komplexität dieser Stoffwechselvorgänge konnten wir wesentliche genetische Faktoren, die mit der Anpassungsfähigkeit der Kühe assoziiert waren, identifizieren. Diese Einflussfaktoren sowie das neu entwickelte Maß für die metabolische Robustheit stellen einen wertvollen Beitrag zum Verständnis des Metabolismus der Milchkuh in der Transitphase dar und können für eine Zucht auf robustere Tiere genutzt werden.

## **CHAPTER 1 GENERAL INTRODUCTION**

The main goal of this thesis is the dissection of the genetic basis of the metabolic adaptation of dairy cows during their early lactation. This was approached from different perspectives including genetic, transcriptomic as well as breeding approaches. In this chapter, we give a short overview of the physiology of dairy cows during their early lactation followed by the introduction of relevant methodologies applied in the different studies.

## **The Transition Period of Dairy Cows**

The transition phase of dairy cows is usually defined as the time frame from three weeks *ante-partum* to three weeks *post-partum* [1]. Within the lactation cycle, this time frame represents the most challenging and critical period for both dairy cows and the dairy farmer industry [2]. Due to the shift from a gestational non-lactating state to the onset of lactation, dairy cows experience abrupt changes in their physiology, metabolism as well as nutrition [3,4]. Insufficient food intake coupled with an increasing energy demand drives dairy cows into an unavoidable state of negative energy balance (NEB). Moreover, due to the selective breeding of high-yielding dairy cows over the last decades, the extent and duration of the NEB during their early lactation has even increased, making them most susceptible to production-related diseases [5,6]. Consequently, a high prevalence of health problems and metabolic and infectious diseases are observed within this time frame [2].

In order to overcome the critical state of NEB in early lactation, dairy cows need to adapt their metabolism. Due to the primary role of glucose for milk production that cannot be replaced by other fuels, alternative energy sources need to be mobilized from body fat and reserves for the maintenance of other vital functions [2]. This leads to the mobilization of non-esterified fatty acids (NEFA) into the bloodstream. The liver, as a key component of the metabolic adaptation, take up large amounts of NEFA to produce glucose, ketone bodies (e.g. beta-hydroxybutyrate, BHBA) and other products. In addition, a number of processes involving numerous regulatory components, metabolites and endocrine factors are

needed to maintain an optimal balance between fuel mobilization, processing and usage [2]. Therefore, a careful coordination and regulation of all the processes driving the metabolic adaptation is required for a successful adaptation.

Even though the main mechanisms of metabolic adaptation are the same for all dairy cows, the variability of how each individual cow deals with this situation is remarkably large [7,8]. Even under the same environmental factors and similar production levels, the ability to cope with the dramatic metabolic changes ranges substantially between cows leading to a hypothesis of a genetic component of the metabolic adaptation [9–11]. If this genetic component is existent, it should manifest itself in the fact that some cows are genetically better suited to cope with the metabolic changes and thus have a better metabolic adaptability than other cows facing the same metabolic load. These cows are then considered metabolically robust. This genetic component is under scrutiny in the present dissertation. To this end, we performed a genome-wide association study (GWAS) in **chapter 2** and a differential gene expression analysis (DGEA) in **chapter 3** in order to identify possible genetic factors influencing the metabolic adaptation. In **chapter 4**, we used a reaction norm sire model to quantify the metabolic robustness and studied its genetic merit and heritability by exploiting relationship information among animals.

## **Genome-wide Association Studies**

The concept of mapping genetic factors to phenotypic traits dates back to the 1990's and is one of the major challenges in animal genetics in order to understand the genetic structure of complex traits [12]. Conventional mapping approaches relying on biparental crosses have led to the detection of numerous quantitative trait loci (QTL) [13]. However, the mapped regions are usually of very low resolutions making it almost impossible to detect causal variants or genes [13]. Recently, the invention of high-throughput genotyping technology has made it possible to survey the whole genome for genetic variants associated

with phenotypic traits. Such studies are referred to as genome-wide association studies (GWAS) and are predominantly designed for the detection of single base-pair changes in the DNA sequence, so called single nucleotide polymorphisms (SNPs). Compared to other genetic variants, such as copy number variations or insertions/deletions, SNPs are the most abundant type of genetic variation distributed across the whole genome. Using SNP genotyping arrays, we are now able to genotype many hundreds of individuals for at least many thousands SNPs at a reasonable cost [13]. Since the introduction of commercial SNP arrays covering around 6K to 777K SNPs for the species cattle a large number of studies have been undertaken for several important traits in dairy cattle including milk yield (e.g. [14,15]), milk quality traits (e.g. [16,17]) and fertility traits (e.g. [18,19]).

Traditionally, GWAS are carried out in a single marker setting, i.e. a statistical test is performed for each genetic marker separately to test whether it has an effect on the phenotype considered. A SNP is regarded as being associated with the phenotype if there is a significant phenotypic difference between individuals carrying different genotypes. Hereby, the method of coding the genotypes determines the underlying model, such as dominant, recessive, multiplicative or additive models [20,21]. Nevertheless, it is common practice to perform GWAS based on an additive model, not least because of its power to also detect dominant effects [21]. For an additive effect model, the genotypes for each individual and SNP are coded as the number of copies of the reference allele, i.e. 0, 1 or 2. The reference allele chosen is usually the allele that occurs less frequently in a given population. An association test is then performed based on a linear regression of the phenotype to the genotype.

In mammals, the total number of SNPs across the whole genome was estimated to be at least many millions [22]. For the species *Bos taurus*, a query on the NCBI dbSNP database (<http://www.ncbi.nlm.nih.gov/snp>) for validated SNPs yielded about 36 million results.

Thus, current SNP arrays can only capture a small fraction of the whole genome, so that causal variants may not be covered. Nevertheless, it is still possible to detect significant associations if the SNP array is dense enough, so that the SNPs are in linkage disequilibrium (LD) with the causal variants [13]. LD refers to the correlation of two loci, i.e. when the joint occurrence of alleles at both loci is more frequent than would be expected by chance. As a general rule, SNPs in close distance to each other are in LD and thus, LD is an important factor which must be accounted for when performing GWAS.

On the one hand, GWAS has opened a number of opportunities for researchers to unveil the genetic structure of complex traits. However, it is also accompanied by a number of challenges for both geneticists and statisticians. From a statistical point of view, a GWAS comprises the testing of many thousands of null hypotheses, simultaneously. Testing each single test with a significance level of  $\alpha = 0.05$  would lead to an inflation of false positives or type I error. Several approaches exist to account for this multiple testing problem. The two most often applied methods are the Bonferroni [23,24] and the false discovery rate (FDR) approach [25].

However, correcting for multiple testing always compromises the statistical power to detect true effects. Therefore, a main disadvantage of GWAS in combination with the single marker analysis (SMA) is the low power resulting from the vast multiple testing problems. Considering the correlation between the tests as a consequence of LD between the SNPs, the problem is even worse, since the most commonly applied methods to correct for multiple testing are based on the assumption of independency tests (e.g. [23–25]). Therefore, especially in human genetics, there has been a major shift of GWAS to multi-marker association studies. In this case, the strategy is to group SNPs according to their biological commonalities, for example SNPs that are in high LD to each other (haplotypes, e.g. [26–30]) or SNPs that are mapped to the same genes (e.g. [31–33]). The main idea of multi-

marker tests is to consider a set of SNPs as one unit, for which a statistical model is established to test the joint effect of all SNPs to the phenotype. By doing so, we are able to reduce the multiple testing problem as well as account for the LD structure of SNPs, at least within SNP-sets [26,34]. Moreover, compared to the detection of single SNPs, the detection of genes is more useful in terms of functional interpretations, since genes and their corresponding proteins are the smallest unit of biological activity [35]. Last but not least, gene-based association studies are more powerful for complex phenotypes that are often affected by multiple loci with small to medium-sized effects [36]. Therefore, in **chapter 2**, we conducted a gene-based association study to identify genes associated with important metabolites in early lactation of the transition cow.

## **RNA Sequencing and Differential Gene Expression Analysis**

The transcriptome is defined as the complete set of all transcripts in a cell. As opposed to the genome, which is static, the transcriptome is a dynamic system, whose composition and quantity is dependent of different factors including development stage or physiological condition of an individual [37]. As an important layer between genotypes and phenotypes, the analysis of transcripts is an important step towards the functional interpretation of genetic studies [38]. One main goal of transcriptome analysis is to quantify the changing gene expression levels. Generally, this quantification is used to compare the gene expression profiles of different conditions and can be carried out for a single individual as well as between different individuals. The method is referred to as differential gene expression analysis (DGE analysis).

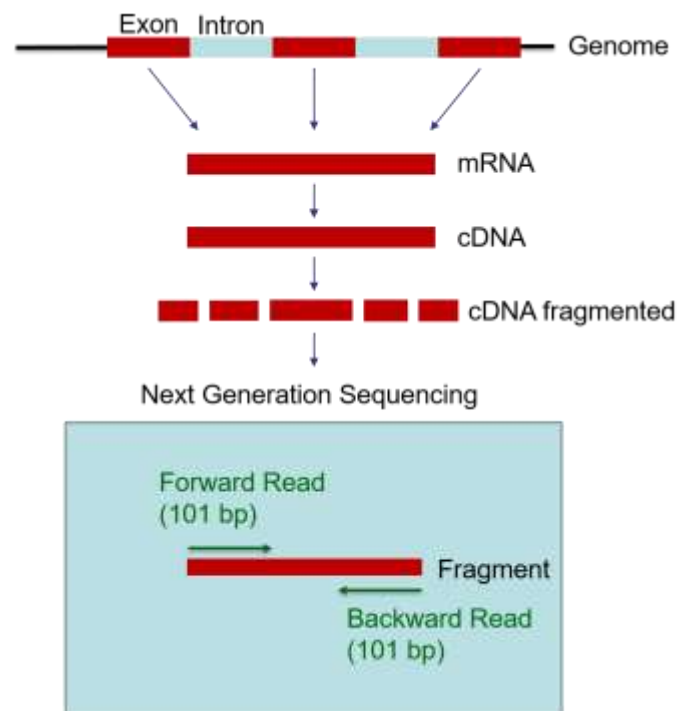
RNA sequencing (RNA-seq) is a relatively new technique that has revolutionized the analysis of DGE analysis. Compared to traditional techniques based on microarrays [37,39], which rely upon existing knowledge about specific gene sequences and have a high level of background noise, RNA-seq sequences the whole transcriptome with high precision to a



single base resolution and is able to detect an almost unlimited range of gene expression changes [37]. In practice, a whole population of RNA representing the expression profile of a specific cell in a certain condition is converted into a library of fragmented cDNA. These fragments are then simultaneously sequenced from either one end (single-end sequencing) or both ends (paired-end sequencing) [37] using next-generation sequencing (high-throughput sequencing) technologies [37,40]. Figure 1.1 presents the schematic principle of RNA-seq for an exemplary gene consisting of three exons and two introns using paired-end sequencing. Depending on the DNA-sequencing technology used, the read length may vary between 30 to 400 base pairs (bp) and each run may produce up to 80 million reads per sample, which is referred to as the transcriptome-wide coverage [41]. However, as opposed to genome sequencing studies, the transcriptome-wide coverage is of less interest owing to the fact that expression levels of different transcripts are highly variable. Consequently, the coverages of different transcripts are also variable and depend strongly on their own expression levels as well as that of other transcripts [38,42]. As a rule, the number of reads that fall into a certain transcript should be linearly proportional to its expression level and this quantity is thus used to measure the gene expression level [37,38]. This measurement of gene expression level is highly accurate as has been shown by validation using qPCR methods [37].

The starting point of the bioinformatic analysis of RNA-seq data are the raw reads for two (or more) sets of samples that we aim to compare. Subsequently, a number of data processing steps have to be conducted, including: (1) quality control of the raw reads, (2) mapping to a reference genome, (3) counting the reads that fall into one transcript using a reference gene annotation. One major challenge of this pipeline is the second step owing to the unique feature of the transcriptome, which is the existence of splice junctions. Numerous tools have been introduced for the alignment of RNA-seq data, for instance STAR [43], Tophat [44], GEM [45], MapSplice [46]. The general strategy is to first conduct a

temporary alignment of the reads in order to discover splice junctions before the final alignment is performed [47]. Further challenges arise when one read is mapped to different transcripts, not mapped at all, or falsely mapped to non-genic regions. The simplest strategy in this situation is to discard all ambiguously mapped reads. Nevertheless, other strategies based on scoring functions or statistical inferences to determine the best possible alignment are also in use [38].



**Figure 1.1:** The principle of RNA-seq illustrated for one gene consisting of three exons and two introns.

The result of the bioinformatic analysis is usually a matrix of read counts for all samples and all annotated genes. However, as mentioned before, the read counts are highly dependent on the sequence coverage (or the total number of reads in a sample) as well as on the length of their corresponding genes. Therefore, a crucial component of RNA-seq analysis is the normalization of the read counts to make them comparable between different samples. An important issue is that the expression of different transcripts may influence each other making the normalization step non-trivial [38]. Nevertheless, several robust methods

exist to correct for this effect [42,48] and as demonstrated on real datasets, the choice of the normalization method does not substantially affect the end results [49]. As for the different gene length, a correction is usually not necessary, since comparisons between different genes are not of interest in DGE analysis.

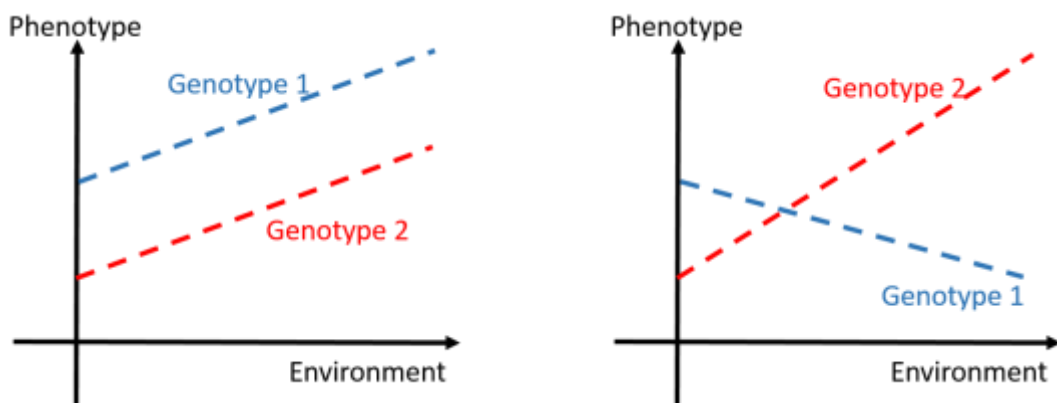
Finally, the normalized read counts are modelled within a statistical framework to test whether and which genes are differentially expressed between different conditions. Due to the nature of the RNA-seq data, the Poisson distribution seems to be the best choice for its approximation, since it is usually used to model count data [38]. One assumption of the Poisson distribution is the equality of its mean and variance. However, as observed from real data analysis, RNA-seq data tends to be over-dispersed [38,48]. Therefore, a common practice for DGE analysis of RNA-seq data is to model the read counts using the negative binomial distribution [38,50]. This approach coupled with a generalized linear model to identify genes differentially expressed before and after lactation, was applied in **chapter 3**.

## **Reaction Norm Models**

The term reaction norm, or norm of reaction, was first introduced by Woltereck in 1909 [51] to describe the genetically determined plasticity of phenotypes. Besides genetic variation, phenotypic plasticity has also long been recognized as an important evolutionary strategy for different species in order to adapt to environment changes [52,53]. The theory of phenotypic plasticity originates from the observation that individuals carrying the same genotypes (clones) may show different expression of the same phenotype when exposed to different environments [51,52]. The change in phenotype as a response to environmental changes is referred to as the phenotypic plasticity. In terms of quantitative traits, the plasticity or equivalently the reaction norm of a phenotype is represented by a function describing the relationship between phenotype and environment. These functions are genotype

specific and depending on whether they are genetically determined or not, they may vary in their shapes for different genotypes [52].

Even though the theory of plasticity is not constrained to any kind of function, in genetics it is common practice to assume a linear relationship between phenotype and environment [52,54]. In this context, the rate of phenotypic change reflected by the slope of the linear function is a measure for the phenotypic plasticity [54]. In Figure 1.2, we present examples of two different linear reaction norm models. In both panels of Figure 1.2, the two dashed lines represent the linear reaction norms of two different genotypes exposed to the same environment. In the left panel, the change in phenotype represented by the slopes of the two curves does not differ between the two genotypes and thus, in this example, the plasticity is not genetically determined. In contrast to the first example, the right panel of Figure 1.2 illustrates the case when the slope of the reaction norm is dependent of the genotype. This kind of effect is also referred to as genotype by environment interaction effect.



**Figure 1.2:** The principle of linear reaction norm models. In both graphs, the two lines depict the linear reaction norm of two different genotypes for a specific phenotype in response to environment change. In the left case, the phenotypic plasticity reflected by the slopes is equal between genotypes and thus, there is no genotype by environment interaction effect. In the right case, the phenotype plasticity is genetically determined, which is reflected by different slopes of the two reaction norms.

In cattle, the conventional and simplest way to model genotype by environment interaction is to consider the same phenotype expressed in different environments as different traits, for which a genetic correlation is estimated [55–57]. However, not least because of globalization, which increases the range of environments that the daughters of sires encounter [58], modelling genotype by environment interaction effects using reaction norm models is becoming a more popular approach in dairy cattle [55,58–61]. This applies a random regression model to model the reaction norms of different genotypes [61]. Such a model usually includes two sets of regressions: (1) a fixed regression consisting of a fixed slope and intercept that is common for all genotypes and (2) a random regression consisting of a random slope and intercept that is specific to each genotype [61]. Here, we purposely use the term genotype instead of individual to clarify the basis concept of reaction norm models. In this context, a genotype is referred to as an individual animal with its specific genetic constellation. However, in order to fit such a model an appropriate number of observations for each individual animal are needed, for which the variable environment is varied. This would mean that the same genotype and thus the same individual has to be exposed to a range of environments, which is practically impossible for the majority of phenotypes. Therefore, in animal breeding, a general strategy in this situation is to exploit relationship structures between animals and to make use of the genetic information shared between related individuals. For instance, in **chapter 4**, we suggested the use of a reaction norm sire model, in which a linear reaction norm is estimated for each bull using the data of his daughters exposed to different challenges and, thus, exhibit different reactions.

## References

1. Drackley JK. Biology of Dairy Cows During the Transition Period: the Final Frontier? J. Dairy Sci. 1999;82:2259–73.
2. Sundrum A. Metabolic Disorders in the Transition Period Indicate that the Dairy Cows' Ability to Adapt is Overstressed. Animals. 2015;5:978–1020.

3. Gross J, van Dorland HA, Bruckmaier RM, Schwarz FJ. Performance and metabolic profile of dairy cows during a lactational and deliberately induced negative energy balance with subsequent realimentation. *J. Dairy Sci.* 2011;94:1820–30.
4. Gross J, van Dorland HA, Schwarz FJ, Bruckmaier RM. Endocrine changes and liver mRNA abundance of somatotrophic axis and insulin system constituents during negative energy balance at different stages of lactation in dairy cows. *J. Dairy Sci.* 2011;94:3484–94.
5. Knaus W. Dairy cows trapped between performance demands and adaptability. *J. Sci. Food Agric.* 2009;89:1107–14.
6. Rauw WM, Kanis E, Noordhuizen-Stassen EN, Grommers FJ. Undesirable side effects of selection for high production efficiency in farm animals: a review. *Livest. Prod. Sci.* 1998;56:15–33.
7. Kessel S, Stroehl M, Meyer HHD, Hiss S, Sauerwein H, Schwarz FJ, et al. Individual variability in physiological adaptation to metabolic stress during early lactation in dairy cows kept under equal conditions. *J. Anim. Sci.* 2008;86:2903–12.
8. van Dorland HA, Richter S, Morel I, Doherr MG, Castro N, Bruckmaier RM. Variation in hepatic regulation of metabolism during the dry period and in early lactation in dairy cows. *J. Dairy Sci.* 2009;92:1924–40.
9. Ingvarlsen KL, Dewhurst RJ, Friggens NC. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. *Livest. Prod. Sci.* 2003;83:277–308.
10. Drackley JK, Dann HM, Douglas GN, Guretzky NAJ, Litherland NB, Underwood JP, et al. Physiological and pathological adaptations in dairy cows that may increase susceptibility to periparturient diseases and disorders. *Growth.* 2005;7.
11. Graber M, Kohler S, Kaufmann T, Doherr MG, Bruckmaier RM, van Dorland HA. A field study on characteristics and diversity of gene expression in the liver of dairy cows during the transition period. *J. Dairy Sci.* 2010;93:5200–15.
12. Falconer DS, Mackay TFC. Introduction to quantitative genetics. 4. ed., [16. print.]. Harlow: Pearson, Prentice Hall; 2009.
13. Zhang H, Wang Z, Wang S, Li H. Progress of genome wide association study in domestic animals. *J. Anim. Sci. Biotechnol.* 2012;3:1.
14. Hayes BJ, Bowman PJ, Chamberlain AJ, Savin K, Tassell CP van, Sonstegard TS, et al. A Validated Genome Wide Association Study to Breed Cattle Adapted to an Environment Altered by Climate Change. *PLOS ONE.* 2009;4:e6676.
15. Jiang L, Liu J, Sun D, Ma P, Ding X, Yu Y, et al. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. *PLoS ONE.* 2010;5:e13661.

16. Bouwman AC, Bovenhuis H, Visker MHPW, van Arendonk JAM. Genome-wide association of milk fatty acids in Dutch dairy cattle. *BMC Genet.* 2011;12:43.
17. Uemoto Y, Abe T, Tameoka N, Hasebe H, Inoue K, Nakajima H, et al. Whole-genome association study for fatty acid composition of oleic acid in Japanese Black cattle. *Anim. Genet.* 2011;42:141–8.
18. Sahana G, Guldbrandtsen B, Bendixen C, Lund MS. Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. *Anim. Genet.* 2010;41:579–88.
19. Huang W, Kirkpatrick BW, Rosa GJM, Khatib H. A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. *Anim. Genet.* 2010;41:570–8.
20. Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief. Bioinform.* 2002;3:146–53.
21. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLOS Comput Biol.* 2012;8:e1002822.
22. Eck SH, Benet-Pagès A, Flisikowski K, Meitinger T, Fries R, Strom TM. Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery. *Genome Biol.* 2009;10:R82.
23. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilità [Internet]. Libreria internazionale Seeber; 1936. Available from: <https://books.google.de/books?id=3CY-HQAACAAJ>
24. Dunn OJ. Estimation of the Medians for Dependent Variables. *Ann. Math. Stat.* 1959;30:192–7.
25. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 1995;57:289–300.
26. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* 2009;33:497–507.
27. Wu Y, Fan H, Wang Y, Zhang L, Gao X, Chen Y, et al. Genome-Wide Association Studies Using Haplotypes and Individual SNPs in Simmental Cattle. *PLOS ONE.* 2014;9:e109330.
28. Hayes BJ, Chamberlain AJ, McPartlan H, Macleod I, Sethuraman L, Goddard ME. Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. *Genet. Res.* 2007;89:215–20.
29. Calus MP, Meuwissen TH, Windig JJ, Knol EF, Schrooten C, Vereijken AL, et al. Effects of the number of markers per haplotype and clustering of haplotypes on the ac-

- curacy of QTL mapping and prediction of genomic breeding values. *Genet. Sel. Evol.* 2009;41:11.
30. Grapes L, Dekkers JCM, Rothschild MF, Fernando RL. Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. *Genetics.* 2004;166:1561–70.
  31. Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, et al. Voxelwise gene-wide association study (vGeneWAS): Multivariate gene-based association testing in 731 elderly subjects. *NeuroImage.* 2011;56:1875–91.
  32. Neale BM, Sham PC. The Future of Association Studies: Gene-Based Analysis and Replication. *Am. J. Hum. Genet.* 2004;75:353–62.
  33. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *Am. J. Hum. Genet.* 2008;82:386–97.
  34. Ha N-T, Gross JJ, van Dorland A, Tetens J, Thaller G, Schlather M, et al. Gene-Based Mapping and Pathway Analysis of Metabolic Traits in Dairy Cows. *PLoS ONE.* 2015;10:e0122325.
  35. Potkin SG, Turner JA, Guffanti G, Lakatos A, Torri F, Keator DB, et al. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. *Cognit. Neuropsychiatry.* 2009;14:391–418.
  36. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* 2009;18:111–7.
  37. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 2009;10:57–63.
  38. Wolf JB. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol. Ecol. Resour.* 2013;13:559–72.
  39. Clark TA, Sugnet CW, Ares M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science.* 2002;296:907–10.
  40. Holt RA, Jones SJM. The new paradigm of flow cell sequencing. *Genome Res.* 2008;18:839–46.
  41. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 2014;15:121–32.
  42. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010;11:R106.
  43. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.



44. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14:R36.
45. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods.* 2012;9:1185–8.
46. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* 2010;38:e178.
47. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods.* 2013;10:1185–91.
48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* 2010;26:139–40.
49. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* 2015;16:59–70.
50. Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* 2012;99:248–56.
51. Woltereck R. Weitere experimentelle untersuchungen über Artänderung, speziell über das Wesen quantitativer Artunterschiede bei Daphniden. *Z. Für Indukt. Abstamm.-Vererbungslehre.* 1913;9:146–146.
52. de Jong G. Quantitative Genetics of reaction norms. *J. Evol. Biol.* 1990;3:447–68.
53. Bradshaw AD. Evolutionary significance of phenotypic plasticity in plants. *Adv. Genet.* 1965;13:115–55.
54. van de Pol M. Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models: Individual variation in reaction norms. *Methods Ecol. Evol.* 2012;3:268–80.
55. Hayes BJ, Carrick M, Bowman P, Goddard ME. Genotype x environment interaction for milk production of daughters of Australian dairy sires from test-day records. *J. Dairy Sci.* 2003;86:3736–44.
56. Veerkamp RF, Goddard ME. Covariance functions across herd production levels for test day records on milk, fat, and protein yields. *J. Dairy Sci.* 1998;81:1690–701.
57. Cromie AR, Kelleher DL, Gordon FJ, Rath M. Genotype by environment interaction for milk yield in dairy cattle. *Ir. Grassl. Anim. Prod. Assoc. J.* 1997;31:29–35.

58. Strandberg E, Brotherstone S, Wall E, Coffey MP. Genotype by environment interaction for first-lactation female fertility traits in UK dairy cattle. *J. Dairy Sci.* 2009;92:3437–46.
59. Kolmodin R, Strandberg E, Madsen P, Jensen J, Jorjani H. Genotype by Environment Interaction in Nordic Dairy Cattle Studied Using Reaction Norms. *Acta Agric. Scand. Sect. — Anim. Sci.* 2002;52:11–24.
60. Streit M, Reinhardt F, Thaller G, Bennewitz J. Genome-wide association analysis to identify genotype  $\times$  environment interaction for milk protein yield and level of somatic cell score as environmental descriptors in German Holsteins. *J. Dairy Sci.* 2013;96:7318–24.
61. Strandberg E, Kolmodin R, Madsen P, Jensen J, Jorjani H. Genotype by environment interaction in Nordic dairy cattle studied by use of reaction norms. *Interbull Bull.* 2000;41.

## CHAPTER 2 GENE-BASED MAPPING AND PATHWAY ANALYSIS OF METABOLIC TRAITS IN DAIRY COWS

*Ngoc-Thuy Ha*<sup>\*1,2</sup>, *Josef Johann Gross*<sup>1</sup>, *Annette van Dorland*<sup>1</sup>, *Jens Tetens*<sup>3</sup>, *Georg Thaller*<sup>3</sup>, *Martin Schlather*<sup>4</sup>, *Rupert Bruckmaier*<sup>1</sup>, *Henner Simianer*<sup>2</sup>

<sup>1</sup> Veterinary Physiology, Vetsuisse Faculty University of Bern, Switzerland

<sup>2</sup> Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University Goettingen, Germany

<sup>3</sup> Institute of Animal Breeding and Husbandry, Christian-Albrechts-University Kiel, Germany

<sup>4</sup> Chair of Mathematical Statistics, University of Mannheim, Germany

\* Email: [ngoc-thuy.ha@vetsuisse.unibe.ch](mailto:ngoc-thuy.ha@vetsuisse.unibe.ch)

Published in PLoS ONE

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0122325>

## **Abstract**

The metabolic adaptation of dairy cows during the transition period has been studied intensively in the last decades. However, until now, only few studies have paid attention to the genetic aspects of this process. Here, we present the results of a gene-based mapping and pathway analysis with the measurements of three key metabolites, (1) non-esterified fatty acids (NEFA), (2) beta-hydroxybutyrate (BHBA) and (3) glucose, characterizing the metabolic adaptability of dairy cows before and after calving. In contrast to the conventional single-marker approach, we identify 99 significant and biologically sensible genes associated with at least one of the considered phenotypes and thus giving evidence for a genetic basis of the metabolic adaptability. Moreover, our results strongly suggest three pathways involved in the metabolism of steroids and lipids are potential candidates for the adaptive regulation of dairy cows in their early lactation. From our perspective, a closer investigation of our findings will lead to a step forward in understanding the variability in the metabolic adaptability of dairy cows in their early lactation.

## **Introduction**

Selective breeding of dairy cows during the last decades has led to the modern high-yielding dairy cow producing more than 45 kg milk per day [1]. However, the immense milk yield also entails a high energy demand during the early lactation period, which cannot be fully covered by feed intake [2]. In order to overcome the metabolic load resulting from a negative energy balance, dairy cows need to mobilize body fat, protein and mineral stores. A failure in metabolic adaptation to this situation leads to an increased susceptibility to health problems as well as development of production-related diseases such as ketosis and fatty liver [3,4].

Numerous studies have tried to elucidate and describe the complex system of metabolic adaptation of dairy cows during their early lactation period [2,5–9]. They identified a num-

ber of crucial regulated target genes, metabolites and endocrine factors in the liver and blood plasma that are involved in important pathways responsible for the adaptive regulation of the metabolism. Nevertheless, so far, a general understanding of why and how the ability of adaptation varies between cows has still not been reached.

Interestingly, even under the same conditions and similar production levels, the success of adaptation differs substantially between cows [5,6]. This strongly suggests that the ability of adaptation may have a genetic basis. In this study, we aim to identify the genetic factors influencing the metabolic adaptation performance during early lactation. In particular, we are interested in identifying genes as well as pathways associated with levels of candidate metabolites in the blood plasma, which were previously confirmed to be essentially involved in the regulation of metabolic adaptation in dairy cows [2,4,6–10].

Genome-wide association studies (GWAS) using single marker analysis (SMA) have been an essential tool for identifying genetic effects. GWAS approaches have been used to detect genomic regions which affect parameters changing during a negative energy balance of dairy cows and milk production-related traits [11–13]. When applied to high density marker data, the SMA approach usually has a massive multiple testing problem, which, when accounted for properly, substantially decreases the power to detect true genetic effects. Another shortcoming of the SMA approach is that it ignores the fact that genes may be represented by several markers and so the effect of a gene may be split up into several marker effects, each of which might not be large enough to pass the significance threshold. Therefore, especially in human genetics, researchers have come forward with gene-based association approaches aiming to overcome the limitations of the SMA [14–19].

In general, the main idea of a gene-based approach is to test each gene instead of each SNP separately by summarizing all SNP-effects annotated to a gene together to one main gene-effect. Here, one main challenge has been how to summarize the SNP-effects reasonably

in order to obtain an efficient gene-level test statistic. To date, a number of methods has been proposed ranging from simple gene-level statistics like the minimum p-value approach [18] to complex statistics, which may account for the linkage disequilibrium (LD) structure or even integrate functional information of the SNPs [16,17,20,21]. For our analysis, we decided to employ the gene-based score test (GBST) adapted from Pan [21]. This approach accounts for the LD structure of the SNPs in a gene and the gene size measured by the number of SNPs, which in turn prevents the inflation of type I error.

In the following, we present the results of the GBST applied to the three key metabolites: (1) non-esterified fatty acids (NEFA), (2) beta-hydroxybutyrate (BHBA) and (3) glucose. We further use the significant genes to detect metabolic pathways potentially affecting these traits to gain understanding of their biological backgrounds. To this end, we adapt the methods commonly used for the analysis of gene expression profiles and gene sets in microarray data experiments, termed gene-set enrichment analysis (GSEA) [22]. More precisely, we employ the permutation-based weighted Kolmogorov Smirnov Test (WKST) by Subramanian et al. [23] and the Wilcoxon Rank Sum Test (WRST) [22] for the identification of pathways, which have been reported to be more successful than other approaches [22,24].

As a result, we found several biologically sensible genes and pathways associated with candidate metabolites during the transition period, which are essential for the adaptation of dairy cows. We further identified the three pathways involved in the metabolism of lipids and steroids, having a joint impact on all of our phenotypes. This may be regarded as evidence for the genetic basis for the adaptation performance of dairy cows and, at the same time, reveals its complexity.

## Results

### *Analysis Overview*

In order to assess the genetic characteristics of the metabolic adaptation of dairy cows during calving and lactation, we examined the three metabolites NEFA, BHBA and glucose at three critical points of time: 3 weeks before calving (T1), 4 and 13 weeks after calving (T2 and T3, respectively). According to van Dorland et al. [6], Graber et al. [10] and Gross et al. [2,4,7], these metabolites are key factors for the metabolic status of dairy cows during their early lactation period.

NEFA and BHBA, both serving as energy sources, are negatively correlated with feed intake and the synthesis of glucose, which is an essential substrate for milk synthesis. In general, dairy cows exhibit increased concentrations of NEFA and BHBA during the transition period resulting in a higher risk for diseases such as ketosis or fatty liver. Hence, we are especially interested in finding genes and pathways that could be responsible for regulating the concentration of NEFA, BHBA and glucose. In particular, we wish to find pathways that are able to inhibit the production of NEFA and BHBA and, at the same time, stimulate the production of glucose during the transition period. To this end, we performed the GBST and GSEA using the measurements of NEFA, BHBA and glucose at T1, T2 and T3. We also considered the changes of the metabolites over time and thus used the ratio of their concentrations measured at the different points of time (T2/T1, T3/T1, and T3/T2) for each metabolite, respectively, as additional phenotypes.

### *Gene-based Analysis and SMA*

For all the three traits considered, the GBST found 99 significant associations with the false discovery rate (FDR) approach [25] and 46 with the Bonferroni-correction at a genome-wide FDR or significance level of 5%. Table 2.1 summarizes the number of significant genes for each of the three phenotypes. A detailed description of the discovered genes

is listed in the supplementary Table 2.1 and the Manhattan plots for all our results can be found in supplementary Figures 2.1-2.3. As a comparison, we also performed a simple SMA that identified only two significant genes (FDR = 5%), which were also detected by the GBST. In the following, all results are based on the GBST at a genome-wide FDR level of 5%.

**Table 2.1:** Number of significant genes with the GBST (FDR = 5%) for the three metabolites.

Time/ratio	NEFA	BHBA	glucose
T1	5	8	5
T2	5	10	3
T3	7	0	12
T2/T1	2	9	8
T3/T1	10	0	3
T3/T2	9	2	1
Sum	38	29	32

First, we concentrate on the analysis of the results for the metabolite NEFA. Here, we found several genes on chromosome 3 and 13 affecting the concentration of NEFA during the ante- (T1) and *post-partum* (T3) period, respectively. However, these genes seem to have no statistical impact on the NEFA concentrations during the early lactation period with the highest metabolic load (T2). Moreover, we discover the gene *SNAI2* (*snail homolog (Drosophila)*) on chromosome 14 to be significantly associated with the ratio of NEFA concentrations measured at T2 and T1 ( $T2/T1$ ,  $p = 6.28 \times 10^{-7}$ ) as well as during T2 ( $p = 5.08 \times 10^{-8}$ ). This strongly supports the view that *SNAI2* is important for the regulation of this metabolite during lactation and hence has an effect on the adaption of dairy cows. Also noticeable are the genes *UGT2B15* (*UDP glucuronosyltransferase 2 family, polypeptide B15*) and *MGC152010* (*UDP glucuronosyltransferase 2 family*) associated with the ratio of NEFA (T3/T1) with p-values of  $p = 1.11 \times 10^{-15}$  and  $p = 1.27 \times 10^{-11}$ , respectively. This in-



icates that the regulation of the metabolic status of dairy cows before and after the onset of lactation is indeed different, even though usually the metabolic load is negligible at these times (T1 and T3).

In a similar fashion, we also identified several significant genes (i.a. *DNAJC30* (*DnaJ* (*Hsp40*) homolog, subfamily C, member 30) and *WBSCR22* (*Williams Beuren syndrome chromosome region 22*)) on chromosome 25 associated with the ratio of BHBA (T2/T1) with p-values smaller than  $1.69 \times 10^{-6}$ . Interestingly, these genes also seem to affect the BHBA concentration strongly (all p-values  $< 9.22 \times 10^{-10}$ ) during the early lactation period with the highest metabolic load (T2) but not 3 weeks before or 13 weeks after calving. Moreover, further investigations demonstrated that cows carrying both a minor allele at the SNP annotated to *DNAJC30* and only major alleles at the three SNPs in *WBSCR22* tend to have higher BHBA concentrations during the transition period (1.53 mmol/L on average) than cows with the opposite genetic characteristics (1.09 mmol/L on average). A Welch two sample t-test comparing the two different groups of cows yielded a p-value of  $p = 0.044$ .

Finally, we focused on the analysis of the phenotype glucose, an essential metabolite for the synthesis of milk during lactation. Here, the gene *UEVLD* (*UEV* and lactate/malate dehydrogenase domains) seems to play an important role in the regulation of the glucose concentration during lactation (T2, T2/T1). Similar to the results of BHBA, we investigated the 15 SNPs annotated to this gene. Among these SNPs, we identified seven markers interacting with each other. Cows that carry only major alleles at these loci have a relatively high concentration of glucose in their blood at T2 (3.3 mmol/L on average, other cows: 3.0 mmol/L on average). Comparing the glucose concentration of these cows with the others yielded a p-value of  $p = 3.3 \times 10^{-7}$  at T2, but only  $p = 0.01$  and  $p = 0.002$  at T1 and T3.

Even though the differences are significant at all three points in time, the effect appears to be highest during the early lactation period (T2, week 4 *post-partum*).

### *Pathway Analysis*

For the pathway analysis, we used the permutation-based weighted Kolmogorov Smirnov Test (WKST) by Subramanian et al. [23] as well as the Wilcoxon Rank Sum Test (WRST) [22] with  $N = 10,000$  permutations. Note that, from a statistical point of view, the tests were performed in a two-step framework (see Methods Section), in which the results of the GBST were used as input information for the WKST and WRST. By doing so, we were able to ignore the uncertainty resulting from the estimation of the p-values for the GBST, which in turn could increase the uncertainty in the estimation of the p-values for the WKST and WRST. In order to account for this, we will only use the empirical p-values obtained by the WKST and WRST in the following analysis to rank the pathways according to their importance, but will not look at their significance. The aim is then to identify biologically and physiologically sensible pathways among the five top ranked pathways with the smallest p-values for each of the three phenotypes.

Supplementary Tables 2.2 and 2.3 show the five top ranked pathways for each of the three phenotypes and points in time with the WKST and WRST, respectively. Due to the huge amount of results, we predominantly concentrated on pathways actively influencing the three phenotypes at T2 and T2/T1. Table 2.2 lists all the pathways ranked at least fifth by both WKST as well as WRST. Among the 20 phenotype-to-pathway associations, we found many associations to be biologically and physiologically sensible. We were further able to connect most of our findings to other studies (see Table 2.2).

On the one hand, according to our expectations, the pathway involved in the synthesis and degradation of ketone bodies is significantly associated with the ratio of the ketone body BHBA as well as the pathway for the metabolism of starch and sucrose with glucose. On

the other hand, we were not able to establish a significant association of the galactose metabolism with glucose during lactation. This association was only significant 13 weeks after calving (T3, compare supplementary Table 2.2).

**Table 2.2:** Phenotype-pathway associations that were at least fifth ranked by the WKST as well as the WRST, and references supporting the corresponding association if known.

Phenotype	Time	Pathways	Literature
NEFA	T2	Histidine metabolism	[26]
		Sulfur metabolism	
	T2/T1	Glycerolipid metabolism	[27]
		Glycerophospholipid metabolism	
BHBA	T2	Taurine and hypotaurine metabolism	[28]
		Retinol metabolism	
		Tyrosine metabolism	
		Inositol phosphate metabolism	
	T2/T1	Steroid hormone biosynthesis	[29]
		Synthesis and degradation of ketone bodies	
		Tryptophan metabolism	
		Inositol phosphate metabolism	
		Steroid biosynthesis	
		Other glycan degradation	
glucose	T2	Fatty acid elongation	[30]
		Ether lipid metabolism	
	T2/T1	Ether lipid metabolism	
		Starch and sucrose metabolism	
		Steroid hormone biosynthesis	[30]
		Glycerophospholipid metabolism	

Finally, we concentrate on the joint analysis of the three metabolites to discover pathways involved in the regulation of all three phenotypes and thus are important for the metabolic adaptation of dairy cows (see Methods Section). Table 2.3 lists the five top ranked path-

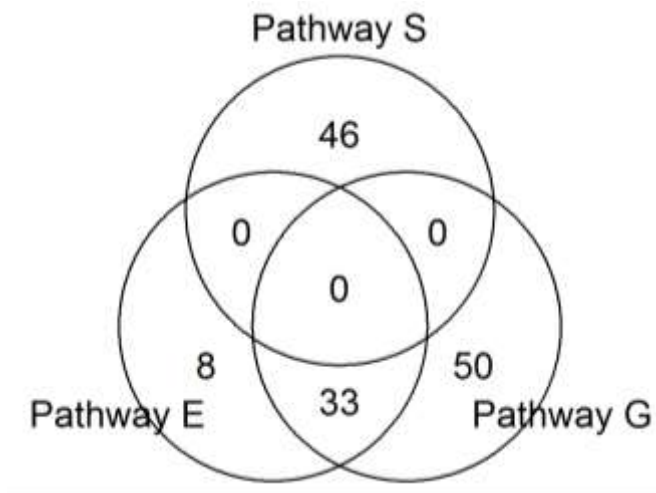
ways resulting from the mentioned analysis (only for T2 and T2/T1, for the complete results please refer to the supplements).

**Table 2.3:** The five pathways with the smallest p-values according to the WKST and WRST for the joined analysis of the NEFA, BHBA and glucose measurements at T2 and T2/T1.

Time	Rank	WKST	P-Value (WKST)	WRST	P-Value (WRST)
<b>T2</b>	<b>1</b>	Steroid hormone biosynthesis	0.0018	Steroid hormone biosynthesis	0.0040
<b>T2</b>	<b>2</b>	Retinol metabolism	0.0063	Other glycan degradation	0.0199
<b>T2</b>	<b>3</b>	Drug metabolism - other enzymes	0.0115	Drug metabolism - cytochrome P450	0.0222
<b>T2</b>	<b>4</b>	Starch and sucrose metabolism	0.0124	Retinol metabolism	0.0223
<b>T2</b>	<b>5</b>	Other glycan degradation	0.0174	Starch and sucrose metabolism	0.0242
<b>T2/T1</b>	<b>1</b>	Ether lipid metabolism	0.0030	Glycerophospholipid metabolism	0
<b>T2/T1</b>	<b>2</b>	Glycerophospholipid metabolism	0.0034	Ether lipid metabolism	5.00E-04
<b>T2/T1</b>	<b>3</b>	Other glycan degradation	0.0061	Nitrogen metabolism	0.0028
<b>T2/T1</b>	<b>4</b>	Tyrosine metabolism	0.0113	Tyrosine metabolism	0.0151
<b>T2/T1</b>	<b>5</b>	Nitrogen metabolism	0.0144	Other glycan degradation	0.0174

Interestingly, we found that the two pathways involved in the metabolism of ether lipids and glycerophospholipids are highly ranked by both the WKST as well as the WRST method. Moreover, both pathways demonstrate empirical p-values smaller than 0.005 at T2/T1. This finding agrees with the results of Klein et al. [31], who were able to establish a link between the glycerophosphocholine levels in milk and the susceptibility for ketosis in dairy cows during early lactation. As for T2, the pathway for steroid hormone biosynthesis is highly associated with the three metabolites, showing p-values of  $p = 0.0018$  and  $p = 0.004$  with the WKST and WRST, respectively. Figure 2.1 illustrates the similarities of the

three pathways and their number of genes. While the two lipid pathways share the majority of the genes involved, they seem to have no similarities with the steroid pathway.



**Figure 2.1:** Venn-diagram for the number of genes annotated to the three pathways steroid hormone biosynthesis (S), ether lipid metabolism (E) and glycerophospholipid metabolism (G).

## Discussion

Metabolic adaptation has been of great interest for dairy scientists during the last decades, but up to the present, little attention has been paid to its genetic aspects. In this study, we investigated the genetic factors influencing the metabolic adaptation of dairy cows during their transition period. In particular, we were interested in the two following questions:

- (1) Do the differences in metabolic adaptation between cows have a genetic basis?
- (2) If there is such a genetic basis, what genes and pathways are responsible for the metabolic adaptation?

As for the first question, our findings strongly support the idea that the metabolic adaption indeed has a genetic basis. Both the gene-based as well as the pathway analysis revealed many genes and pathways influencing the three metabolites, but only at certain points of

time. For instance, the gene *UEVLD* appears to affect the phenotype glucose only in the early lactation, but not 4 weeks before or 13 weeks after calving. The opposite case can be observed from the relations between several genes on chromosome 3 and 13 and the phenotype NEFA.

With regard to the second question, we found several significant genes and pathways regulating the concentrations of NEFA, BHBA and glucose during the transition period. Three pathways with a number of SNPs detected (steroid hormone biosynthesis, ether lipid metabolism and glycerophospholipid metabolism) were found to jointly affect the key metabolites NEFA, BHBA and glucose. The genes distributing to the significance of the three pathways are involved in various sectors of the lipid metabolism. Especially, the repeated link to the cholesterol metabolism in dairy cows coping with elevated NEFA concentrations is obvious and was recently shown at a physiological level [32]. Interestingly, no SNPs were detected for genes, which are directly involved in the ketogenesis. However, the findings regarding associations to different phenotypes in BHBA concentrations are very probably an indirect result of the changes in NEFA concentrations. The availability or surplus of NEFA is a main factor determining the degree of BHBA synthesis, e.g. ketogenesis is regulated through NEFA plasma concentration. Similar to BHBA, glucose concentration cannot be directly connected to significant associations with SNPs in the presented genes and pathways. However, the occurrence of reduced glucose availability at high BHBA concentrations was recently demonstrated [33] and is presumably the reason for the low glucose concentration as a secondary effect of high NEFA concentration.

Generally, we were able to connect most of our findings to previous studies and hence confirm their plausibility. As an example, the associations of the glycerolipid and glycerophospholipid metabolism with the ratio of NEFA concentrations at T2 and T1 may be explained by the intense mobilization of lipids from tissue stores in the transition period

resulting in a substantial release of NEFA into the blood stream [27]. Another interesting pathway associated with NEFA is the histidine metabolism (at T2). While there is no apparent connection between histidine and NEFA, Vanhatalo et al. [26] found that histidine significantly increases the milk and milk protein yield during lactation, but at the same time decreases its lactose and fat contents. Nevertheless, when interpreting these results, we have to take into account that even though the three considered metabolites are indeed indicators for the metabolic adaptation of dairy cows, they may not be fully indicative to the whole process of adaptation. Therefore, we suggest that the discovered genes and pathways should be viewed as potential candidates for closer investigation and validation from a biological perspective in future studies considering the complex endocrine and metabolic interactions.

**Table 2.4:** Comparison of the per test significance threshold between the SMA and GBST after adjustment with the Bonferroni-correction at a genome-wide level  $\alpha = 0.05$ .

	$m$ (# of SNPs/genes)	$\alpha/m$	$-\log_{10}(\alpha/m)$
SMA	601,455 SNPs (all)	$8.31 \times 10^{-8}$	7.08
	231,712 SNPs (intragenic)	$2.16 \times 10^{-7}$	6.67
GBST	22,025 genes	$2.27 \times 10^{-6}$	5.64

In order to answer the two mentioned questions, we performed a GWAS using the GBST as suggested by Pan [21] followed by a GSEA for the identification of pathways. Even though the focus of this work was not methodological, our results demonstrated that using the GBST is more successful than the conventional SMA in identifying biologically sensible genes even with a relatively small samples size. For all of the considered traits, we discovered highly significant genes consisting of many SNPs interacting with each other. The SMA, however, missed most of these genes, since it is only designed to dissect single SNP-effects and has low power due to the massive multiple testing problem. Using the

example of the Bonferroni-correction, Table 2.4 explains the apparent loss of power of the SMA compared to the GBST. By testing each single SNP separately, the SMA needs to test about 30 times more null hypotheses than the GBST resulting in much more conservative significance threshold per test.

The gene-based approach in combination with the pathway analysis is a well-established and commonly used method in human genetics. Here, a number of methods have been proposed and successfully employed to identify genes and pathways contributing to the development of complex human diseases (for some examples, see [15,16,19,20,34]). In animal genetics and breeding, researchers in general are still relying on simple SNP-based association studies. However, Peng et al. [15] as well as our study have shown that complex phenotypes are often affected by the joint action of many variants within a gene or even of many genes within a pathway. Hence, we believe that the use of the gene- and pathway-based approach in animal science is a promising tool to shed new light on the genetic complexity of common traits and deepen the understanding of their biological backgrounds.

## **Material and Methods**

### *Phenotype Data*

The data used in our analyses were obtained from a large on-farm study involving 232 dairy multiparous cows from different breed types (Brown Swiss, Holstein, Swiss Fleckvieh) housed at 64 farms (at least 2 cows per farm used in the trial, calving between November 2007 and April 2008, similar diet: grass and maize silage based feeding with additional concentrate, all cows under supervision of breeding associations) [35] and 50 Holstein dairy cows kept under controlled feeding conditions on an experimental farm (grass and maize silage based feeding with additional concentrate) [2]. In brief, blood samples were taken from cows which had a significant metabolic load in their previous early lactation as estimated by the fat:protein-ratio and milk fat content reflecting a tremendous body



fat mobilization. Based on previous frequent measurements during the transition period, samples were taken at three critical stages of lactation: T1 = week 3 before expected calving (not lactating and no metabolic load); T2 = week 4 *post-partum* (lactating and high metabolic load), and T3 = week 13 after parturition (lactating and no metabolic load). Plasma concentrations of NEFA, BHBA and glucose were measured at T1, T2 and T3 using commercial kits as described by Graber et al. [35] and Gross et al. [2]. Table 2.5 presents the data summary for the three phenotypes and the supplement Figures 2.4-2.6 the corresponding histograms.

**Table 2.5:** Mean and standard deviation of the three metabolites NEFA, BHBA and glucose.

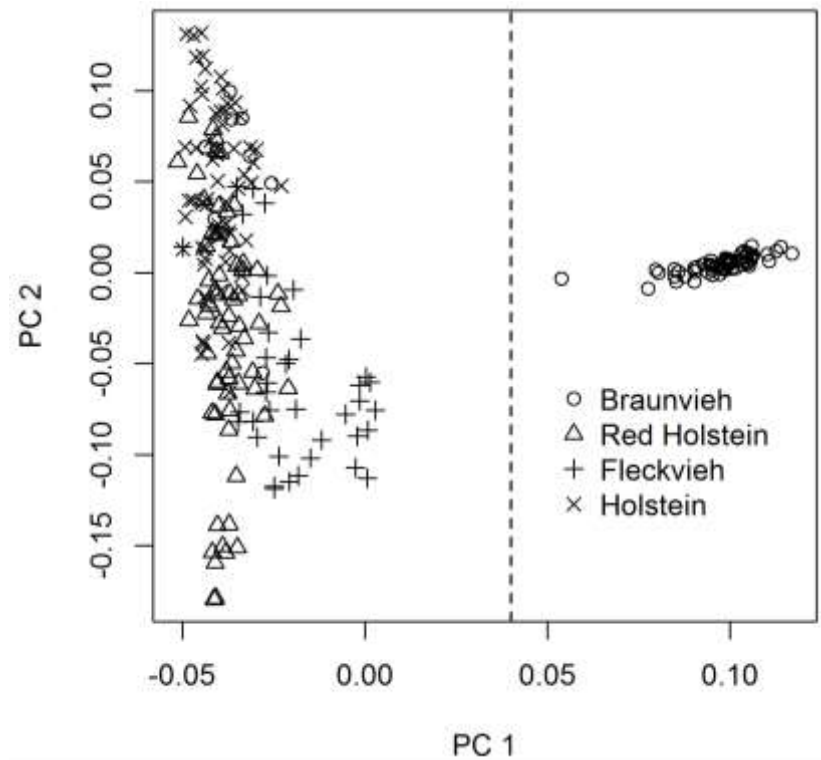
$\mu \pm \sigma$	NEFA log(mmol/L)	BHBA mmol/L	Glucose mmol/L
T1	$4.269 \pm 0.691$	$0.582 \pm 0.266$	$3.719 \pm 0.360$
T2	$5.624 \pm 0.647$	$1.349 \pm 1.081$	$3.146 \pm 0.560$
T3	$4.643 \pm 0.642$	$0.770 \pm 0.352$	$3.702 \pm 0.433$

### *Genotype Data*

The 282 dairy cows were genotyped with the Illumina next-generation High-Density Bovine BeadChip. The resulting genotype dataset, consisting of 777,692 markers for 282 dairy cows, was then quality controlled and filtered with a SNP call rate of 95% and minor allele frequency (MAF) of 5%. After filtering and quality control, 601,455 SNPs for 282 animals remained.

To assess a possible stratification of the material, a principal component analysis (PCA) based on the genotypes was performed. Figure 2.2 displays the first two principal components of the PCA. According to the results, we divided the 282 animals into two groups (Holstein, Red Holstein and Fleckvieh vs. Braunvieh). This grouping reflects that Swiss

Fleckvieh historically was heavily interbred with Red Holstein. Within the two subgroups missing genotypes were imputed with the program BEAGLE (Version 3.3.2 [36]). After removing 38 cows with missing phenotypes, a final dataset with 178 animals in group 1 (**G1**) and 66 animals in group 2 (**G2**) was available for our analyses.



**Figure 2.2:** The two leading principal components of the analysis with the genotype data with 601,455 SNPs for 282 animals. The first and second principal components explain 8.7 % and 1.4 % of the total variation, respectively. The dashed line indicates the division of the cows into two groups.

### *Gene and Pathway Annotation*

In order to perform the GBST and pathway analysis, we need a gene-annotation that allocates SNPs to genes and a pathway-annotation for genes to pathways. For the gene-annotation, we downloaded a list of all known genes from the Ensembl Genes database (Release 73, UMD3.1) for the species *Bos taurus* [37]. We allocated SNPs to genes according to the transcription-start and -end positions including exon, intron, UTR variants as well as SNPs up to 5kbp up- and downstream. The final gene-annotation includes

22,025 genes containing at least one intragenic SNP. For the pathway analysis, we retrieved a gene to pathway annotation from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [29]. We further matched the different gene identifiers from the Ensembl and KEGG database to obtain a final pathway-annotation for our dataset with 81 metabolic pathways containing at least five genes.

### *Statistical Model*

Denote  $y$  the observed vector of phenotypes of  $n$  individuals and  $X$  an  $n \times m$  matrix with values in  $\{0, 1, 2\}$  representing the genotypes for a gene  $G$  consisting of  $m$  SNPs. To test the association of the gene  $G$  to the phenotype  $y$ , we fitted a linear regression model

$$y = Z\theta + X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n),$$

where  $Z$  denotes the  $n \times (k + 1)$  matrix of covariates accounting for possible environmental factors or population structures, including the intercept and  $\theta = (\theta_1, \theta_2, \dots, \theta_{k+1})^t$ ,  $\beta = (\beta_1, \beta_2, \dots, \beta_m)^t$  are the regression coefficients. Accordingly, the statistical problem is to test whether the phenotype  $y$  is influenced by any of the  $m$  SNPs, that is  $H_0: \beta = 0$ . For the SMA approach, we performed a single marker regression and employed a simple t-test based on the same statistical model, but only for one SNP ( $m = 1$ ) at a time.

### *Population Stratification and Environmental Factors*

The principal component analysis based on the genotypes (see Figure 2.2) revealed a substantial population structure within the group **G1** with 178 cows from the breeds Holstein, Red Holstein and Fleckvieh. To avoid inflation of the type I error [38], it is important to account for this stratification both in the gene-based as well as single marker analysis. We conducted a PCA for the group **G1** as to obtain covariates for the regression model and employed the Tracy-Widom test [38] to assess how many of the leading principal components contributed significantly to the population structure. As a result, the first 17 principal

components were significant ( $\alpha = 0.01$ ) and thus were added as covariates to the linear regression model for the group **G1**. We chose a very conservative significant threshold in order to avoid overcorrection of the regression model, which would in turn mask possible true effects. The covariate matrix for the whole dataset (**G1+G2**) with  $n = 244$  animals is then given by  $Z = (1_n, I_{G1}, PC_1, PC_2, \dots, PC_{17})$ , where  $1_n = (1, 1, \dots, 1)^t$ ,  $(I_{G1})_j = 1$  if animal  $j$  is in Group **G1** and 0 otherwise, and  $(PC_i)_j$  is the value of the  $i$ -th principal component for animal  $j$  if it belongs to group **G1** and 0 otherwise for  $j = 1, 2, \dots, n$ .

The NEFA, BHBA and glucose measurements of the 244 cows are from two different studies, which might have different management systems. In order to account for this, we include an additional covariate  $I_E$  with  $(I_E)_j = 1$  if animal  $j$  is from the on-farm study and 0 otherwise. For comparison purposes, we perform the gene-based and pathway analysis twice, with the covariate matrix  $Z$  and  $Z^* = (1_n, I_E, I_{G1}, PC_1, PC_2, \dots, PC_{17})$ . The results of the two analyses are similar, however, the analysis with the matrix  $Z$  yields smaller p-values. Therefore, in this study, we will present results based on the matrix  $Z$ .

### *Gene-based Score Test*

For our gene-based analysis, we employed a modification of the Score Test, motivated by Pan [21]. To this end, denote  $U$  the score vector for the  $m$  SNPs adjusted for the covariates given by

$$U = \frac{1}{\hat{\sigma}} X^t (y - P_Z y),$$

where  $P_Z = Z(Z^t Z)^{-1} Z^t$  and  $\hat{\sigma}$  is the Maximum-Likelihood estimate for  $\sigma$  under  $H_0$ . Then, the covariance matrix of  $U$ , also adjusted for the covariates  $Z$ , is

$$C = (X - P_Z X)^t (X - P_Z X).$$

To test  $H_0$ , Pan [21] suggests using the test statistic

$$SSU = U^t \text{Diag}(C)^{-1} U,$$

where  $\text{Diag}(C)$  is a diagonal matrix with only the diagonal elements of  $C$  as non-zero entries, instead of the standard score statistic  $SSU = U^t C^{-1} U$ . To calculate the distribution of the score statistic  $SSU$ , we used a  $\chi^2$ -approximation by Zhang [39], while accounting for the fact that  $\sigma^2$  was estimated by  $\sigma^2 = \frac{1}{n}(y - P_Z y)^t (y - P_Z y)$ .

According to Pan [21], the modified test statistic  $SSU$  is more powerful than  $SSU'$  for simulated genetic data. Freytag and Bickeboeller [40] also confirmed this finding by employing this method in the context of gene ranking. Using a simulation study, they compared this method to other summary statistics for genes and reported that the test based on the score statistic outperforms all other methods in many scenarios, in particular, when interactions between SNPs are present.

### *Pathway Analysis*

Using the p-values obtained by the GBST, we performed a gene set enrichment analysis (GSEA) to identify metabolic pathways associated with the phenotypes. For this purpose, we adapted the weighted Kolmogorov-Smirnov Test (WKST) suggested by Subramanian et al. [23]. This method was shown to be superior compared to other gene set-level statistics by Hung et al. [22].

Denote  $S$  a given set of genes (e.g. a pathway) and  $L = \{g_1, g_2, \dots, g_N\}$  a ranked list of all genes according to a ranking metric  $r(g_j) = r_j$  with  $r_1 \geq r_2 \geq \dots \geq r_N$ . Importantly, the metric  $r$  should reflect the ‘importance’ of a gene for the phenotype under consideration. In order to test the association of the pathway  $S$  and the considered phenotype according to the GSEA approach, we first calculate an Enrichment Score  $ES$  for this pathway and then permute the phenotypes to obtain its null distribution. To this end, we start with a running-sum  $RS = 0$ . We then walk down the list  $L$  ( $i = 1, 2, \dots, N$ ) and increase  $RS$  by

$$\frac{r_i}{\sum_{g_j \in S} r_j}$$

if the gene  $g_i$  is contained in the pathway  $S$ , and decrease  $RS$  by

$$\frac{1}{N - N_S}$$

if the considered gene is not contained in the pathway  $S$ , where  $N_S = |S|$  is the number of genes in the pathway  $S$ . Finally, the Enrichment Score  $ES$  is defined by the maximum deviation of the running sum  $RS$  from zero. To assess the significance of  $ES$  while accounting for the correlation structure of the genes, we permuted the phenotypes and repeated the whole procedure 10,000 times for each pathway and phenotype. For comparison purposes, we also employed the permutation-based Wilcoxon Rank Sum Test (WRST) [22] based on the test statistic

$$RS = \sum_{g_j \in S} Rank_L(g_j)$$

In our analyses, we ranked our genes according to the p-values obtained by the GBST and set  $r(g_j) = -\log_{10}(p_j)$ . The analysis was performed for each metabolite and each points of time separately. Furthermore, in order to investigate whether there is a pathway that is able to regulate all three metabolites simultaneously, we also conducted a joint pathway analysis for the three metabolites. Here, we ranked the genes according to the product of the p-values obtained from the gene-based analysis with the three different phenotypes and set  $r(g_j) = -\log_{10}(p_j^{NEFA} p_j^{BHBA} p_j^{glucose})$ . This will especially raise the rank of genes exhibiting small p-values for all of the three metabolites. As a result, we hope to discover pathways involving in the regulation of all the three phenotypes and thus are important for the metabolic adaptation of dairy cows.

### *Multiple Hypothesis Testing*

Both the SMA and the GBST require the testing of multiple hypotheses (SNPs or genes) simultaneously. Therefore, it is necessary to adjust the significance threshold of each single test properly in order to avoid an inflation of the genome-wide type I error rate. A commonly used, but very conservative method is the Bonferroni-correction. Here, the significance threshold for each single test  $\alpha'$  is given by the genome-wide rate  $\alpha$  (e.g.  $\alpha = 0.05$ ) divided by the number of hypotheses tested  $m$  ( $\alpha' = \frac{\alpha}{m}$ ). By doing so, the probability for detecting at least one false positive signal, also termed as the familywise error rate (FWER), has an upper boundary determined by the level  $\alpha$ . Another and more powerful method to control the type I error rate is the False Discovery Rate (FDR) approach by Benjamini and Hochberg [25]. As opposed to the Bonferroni-correction, the FDR approach aims to keep the expected proportion of false positives instead of the FWER below a certain level  $q$ . Thereby, the per test significance level  $\alpha'$  is determined so that the equation  $\frac{\alpha' m}{R(\alpha')} \leq q$  holds, where  $R(\alpha')$  denotes the number of tests that were declared significant at the level  $\alpha'$ .

### **Acknowledgement**

This study was supported by Swissgenetics (Zollikofen, Switzerland) and parts of the research have been conducted within the research training group ‘Scaling Problems in Statistics’ (RTG 1644) supported by the Deutsche Forschungsgemeinschaft (DFG).

### **Supplementary Tables and Figures**

**Supplementary Table 2.1:** Results of the GBST. Description of all significant genes

(FDR < 5%) resulting from the gene-based score test (GBST) according to the Ensembl database.

(<http://journals.plos.org/plosone/article/asset?unique&id=info:>

[doi/10.1371/journal.pone.0122325.s007](https://doi.org/10.1371/journal.pone.0122325.s007))

doi:10.1371/journal.pone.0122325.s007

**Supplementary Table 2.2:** Results of the WKST. The five top ranked pathways according to the results of the weighted Kolmogorov Smirnov test (WKST).

(<http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0122325.s008>)

doi:10.1371/journal.pone.0122325.s008

**Supplementary Table 2.3:** Results of the WRST. The five top ranked pathways according to the results of the Wilcoxon Rank Sum test (WRST).

(<http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0122325.s009>)

doi:10.1371/journal.pone.0122325.s009

**Supplementary Table 2.4:** Results of the joint analysis with the WKST. The ten top ranked pathways according to the results of the joint analysis with all three metabolites with the weighted Kolmogorov Smirnov test (WKST).

(<http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0122325.s010>)

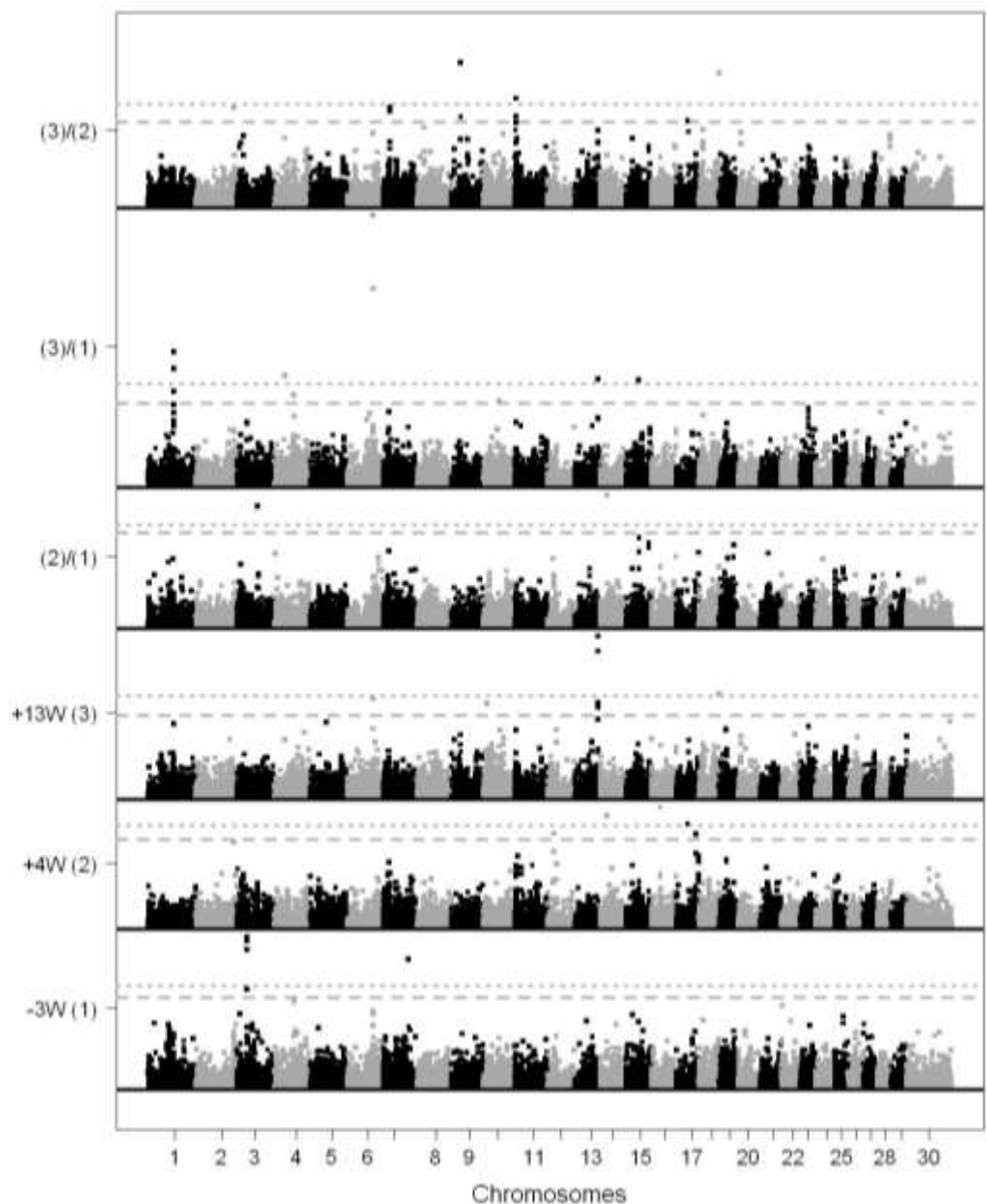
doi:10.1371/journal.pone.0122325.s010

**Supplementary Table 2.5:** Results of the joint analysis with the WRST. The ten top ranked pathways according to the results of the joint analysis with all three metabolites with the Wilcoxon Rank Sum test (WRST).

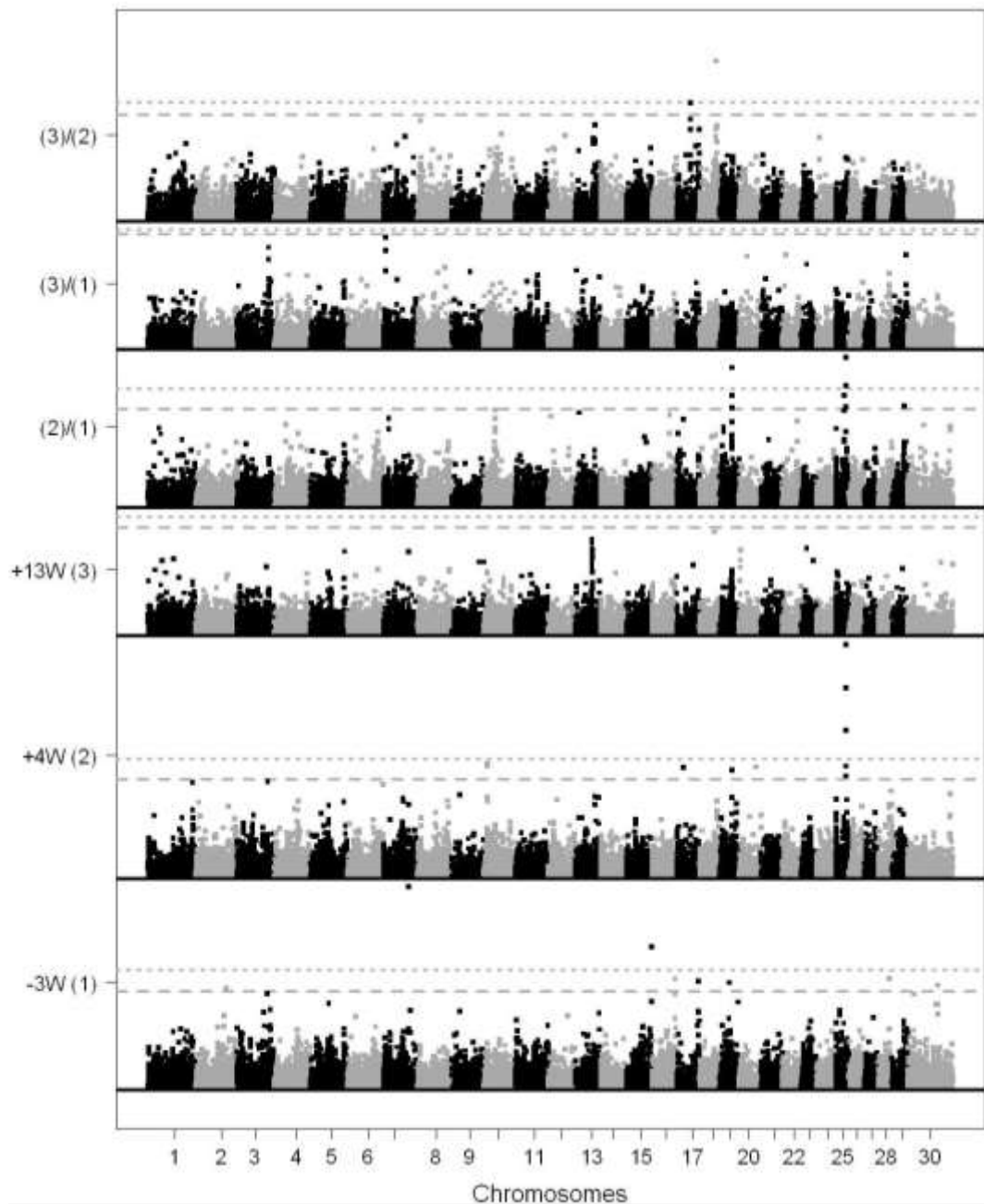
(<http://journals.plos.org/plosone/article/asset?unique&id=info:doi/10.1371/journal.pone.0122325.s011>)

doi:10.1371/journal.pone.0122325.s011

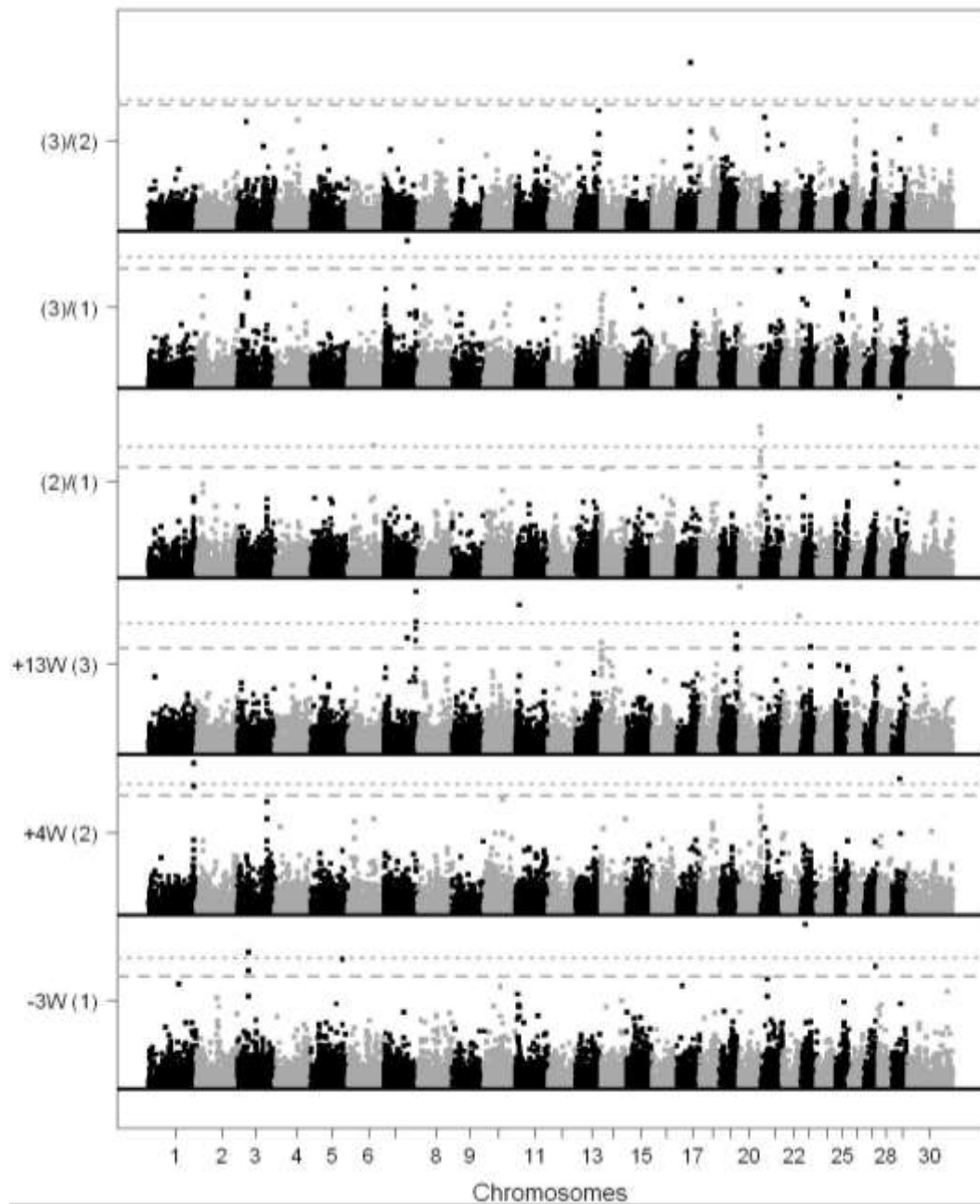




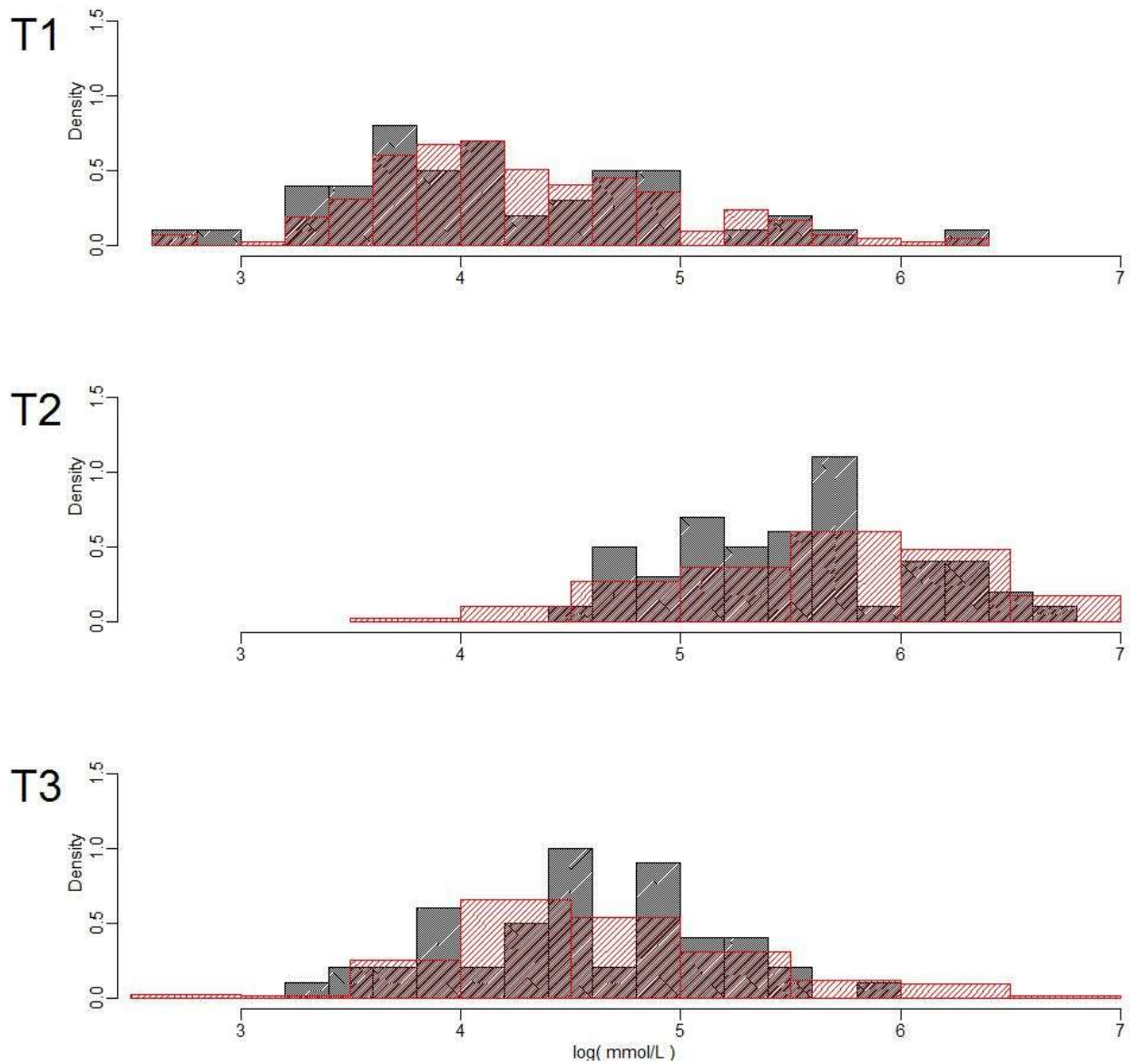
**Supplementary Figure 2.1:** Results of the GBST for NEFA. Manhattan plot of the GBST for the phenotype NEFA measured at T1 (1), T2 (2) and T3 (3) as well as the ratios. Each dot represents a gene. The dotted and dashed lines show the significance thresholds after the multiple testing correction according to Bonferroni and the FDR methods, respectively.



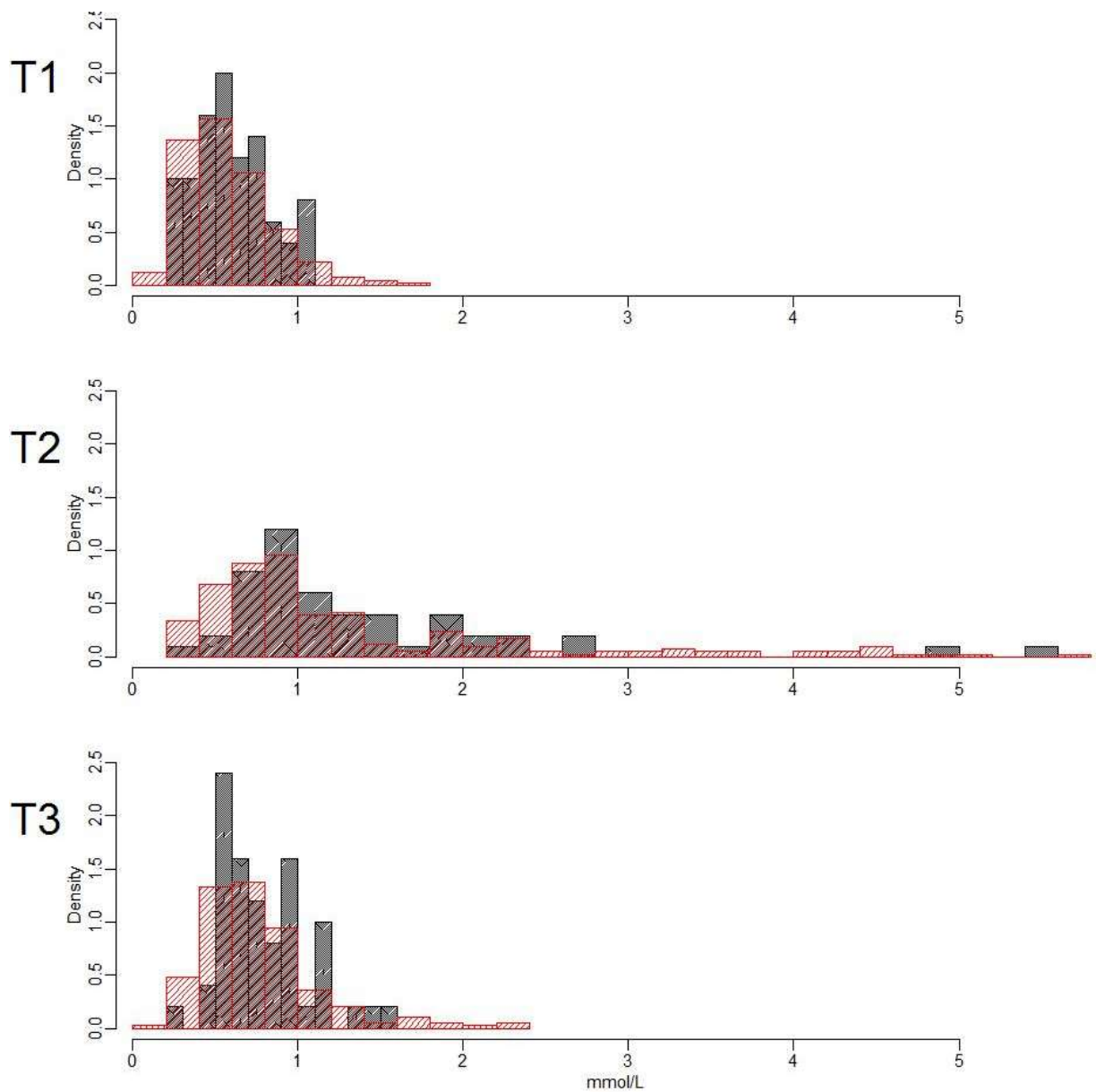
**Supplementary Figure 2.2:** Results of the GBST for BHBA. Manhattan plot of the GBST for the phenotype BHBA measured at T1 (1), T2 (2) and T3 (3) as well as the ratios. Each dot represents a gene. The dotted and dashed lines show the significance thresholds after the multiple testing correction according to Bonferroni and the FDR methods, respectively.



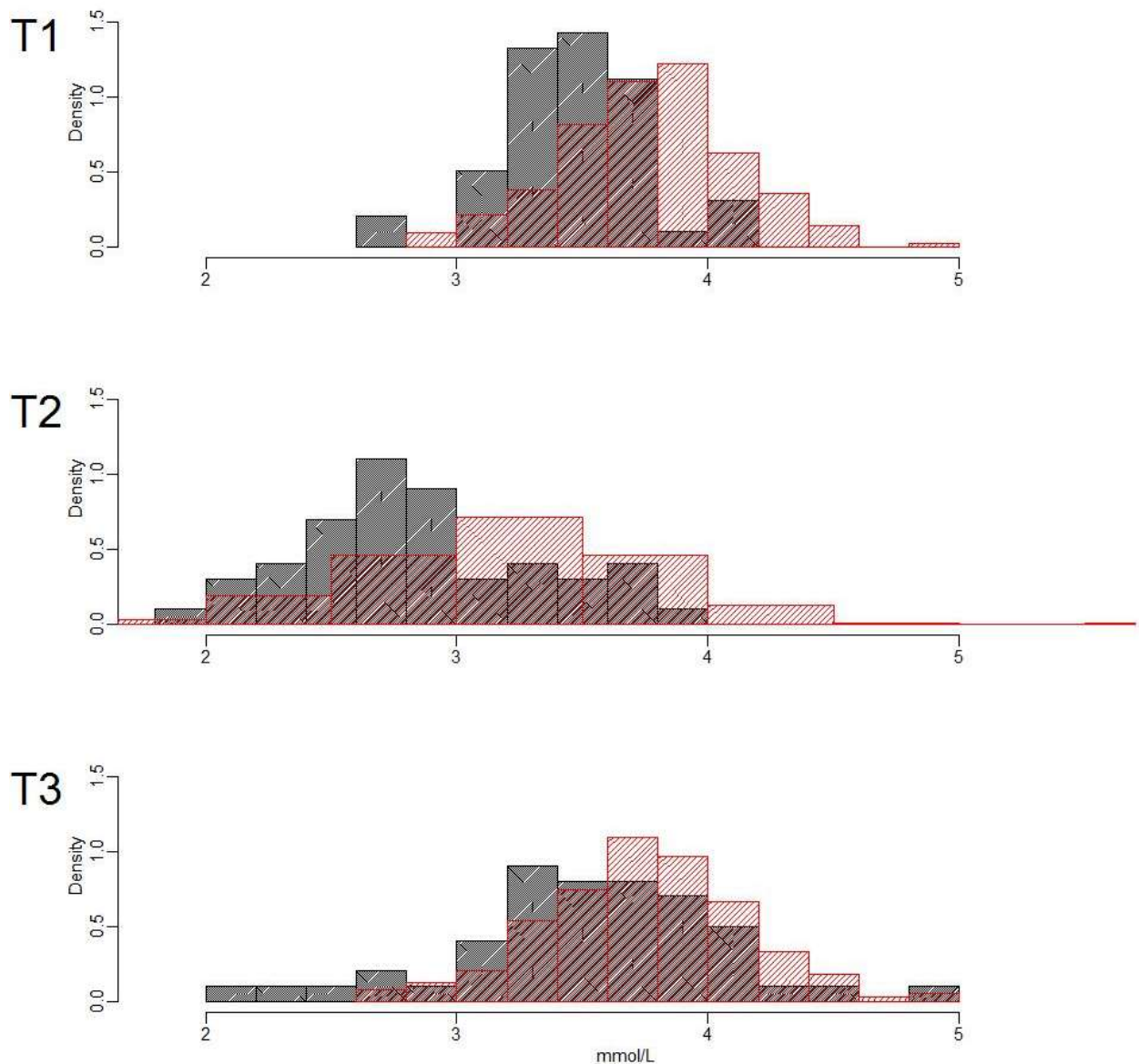
**Supplementary Figure 2.3:** Results of the GBST for glucose. Manhattan plot of the GBST for the phenotype glucose measured at T1 (1), T2 (2) and T3 (3) as well as the ratios. Each dot represents a gene. The dotted and dashed lines show the significance thresholds after the multiple testing correction according to Bonferroni and the FDR methods, respectively.



**Supplementary Figure 2.4:** Histograms of NEFA. Histograms of the phenotype NEFA measured at T1, T2 and T3. The different colors indicate the two different studies (grey = on-farm study).



**Supplementary Figure 2.5:** Histograms of BHBA. Histograms of the phenotype BHBA measured at T1, T2 and T3. The different colors indicate the two different studies (grey = on-farm study).



**Supplementary Figure 2.6:** Histograms of glucose. Histograms of the phenotype glucose measured at T1, T2 and T3. The different colors indicate the two different studies (grey = on-farm study).

## References

1. Wellnitz O, Bruckmaier R, Blum J. Milk ejection and milk removal of single quarters in high yielding dairy cows. *Milk Sci. Int.* 1999;54.
2. Gross J, van Dorland HA, Bruckmaier RM, Schwarz FJ. Performance and metabolic profile of dairy cows during a lactational and deliberately induced negative energy balance with subsequent realimentation. *J. Dairy Sci.* 2011;94:1820–30.
3. Bobe G, Young JW, Beitz DC. Invited review: pathology, etiology, prevention, and treatment of fatty liver in dairy cows. *J. Dairy Sci.* 2004;87:3105–24.

4. Gross JJ, Schwarz FJ, Eder K, van Dorland HA, Bruckmaier RM. Liver fat content and lipid metabolism in dairy cows during early lactation and during a mid-lactation feed restriction. *J. Dairy Sci.* 2013;96:5008–17.
5. Kessel S, Stroehl M, Meyer HHD, Hiss S, Sauerwein H, Schwarz FJ, et al. Individual variability in physiological adaptation to metabolic stress during early lactation in dairy cows kept under equal conditions. *J. Anim. Sci.* 2008;86:2903–12.
6. van Dorland HA, Richter S, Morel I, Doherr MG, Castro N, Bruckmaier RM. Variation in hepatic regulation of metabolism during the dry period and in early lactation in dairy cows. *J. Dairy Sci.* 2009;92:1924–40.
7. Gross J, van Dorland HA, Schwarz FJ, Bruckmaier RM. Endocrine changes and liver mRNA abundance of somatotrophic axis and insulin system constituents during negative energy balance at different stages of lactation in dairy cows. *J. Dairy Sci.* 2011;94:3484–94.
8. Drackley JK. ADSA Foundation Scholar Award. Biology of dairy cows during the transition period: the final frontier? *J. Dairy Sci.* 1999;82:2259–73.
9. Weber C, Hametner C, Tuchscherer A, Losand B, Kanitz E, Otten W, et al. Hepatic gene expression involved in glucose and lipid metabolism in transition cows: Effects of fat mobilization during early lactation in relation to milk performance and metabolic changes. *J. Dairy Sci.* 2013;96:5670–81.
10. Graber M, Kohler S, Müller A, Burgermeister K, Kaufmann T, Bruckmaier RM, et al. Identification of plasma and hepatic parameters related to metabolic robustness in dairy cows. *J. Anim. Physiol. Anim. Nutr.* 2012;96:75–84.
11. Tetens J, Seidenspinner T, Buttchereit N, Thaller G. Whole-genome association study for energy balance and fat/protein ratio in German Holstein bull dams. *Anim. Genet.* 2013;44:1–8.
12. Jiang L, Liu J, Sun D, Ma P, Ding X, Yu Y, et al. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. *PLoS ONE.* 2010;5:e13661.
13. Mai MD, Sahana G, Christiansen FB, Guldbrandtsen B. A genome-wide association study for milk production traits in Danish Jersey cattle using a 50K single nucleotide polymorphism chip. *J. Anim. Sci.* 2010;88:3522–8.
14. Neale BM, Sham PC. The Future of Association Studies: Gene-Based Analysis and Replication. *Am. J. Hum. Genet.* 2004;75:353–62.
15. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* 2009;18:111–7.

16. Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, et al. A Versatile Gene-Based Test for Genome-wide Association Studies. *Am. J. Hum. Genet.* 2010;87:139–45.
17. Li M, Wang K, Grant SFA, Hakonarson H, Li C. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics.* 2009;25:497–503.
18. Kang G, Jiang B, Cui Y. Gene-based Genomewide Association Analysis: A Comparison Study. *Curr. Genomics.* 2013;14:250–5.
19. Huang H, Chanda P, Alonso A, Bader JS, Arking DE. Gene-Based Tests of Association. *PLoS Genet.* 2011;7:e1002177.
20. Li M-X, Gui H-S, Kwan JSH, Sham PC. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *Am. J. Hum. Genet.* 2011;88:283–93.
21. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* 2009;33:497–507.
22. Hung J-H, Yang T-H, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.* 2011;13:281–91.
23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102:15545–50.
24. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics.* 2009;10:47.
25. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 1995;57:289–300.
26. Vanhatalo A, Huhtanen P, Toivonen V, Varvikko T. Response of dairy cows fed grass silage diets to abomasal infusions of histidine alone or in combinations with methionine and lysine. *J. Dairy Sci.* 1999;82:2674–85.
27. Contreras GA, Sordillo LM. Lipid mobilization and inflammatory responses during the transition period of dairy cows. *Comp. Immunol. Microbiol. Infect. Dis.* 2011;34:281–9.
28. LeBlanc SJ, Herdt TH, Seymour WM, Duffield TF, Leslie KE. Peripartum serum vitamin E, retinol, and beta-carotene in dairy cattle and their associations with disease. *J. Dairy Sci.* 2004;87:609–19.



29. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012;40:D109–14.
30. Marks PA, Banks J. INHIBITION OF MAMMALIAN GLUCOSE-6-PHOSPHATE DEHYDROGENASE BY STEROIDS\*. *Proc. Natl. Acad. Sci. U. S. A.* 1960;46:447–52.
31. Klein MS, Buttchereit N, Miemczyk SP, Immervoll A-K, Louis C, Wiedemann S, et al. NMR metabolomic analysis of dairy cows reveals milk glycerophosphocholine to phosphocholine ratio as prognostic biomarker for risk of ketosis. *J. Proteome Res.* 2012;11:1373–81.
32. Kessler EC, Gross JJ, Bruckmaier RM, Albrecht C. Cholesterol metabolism, transport, and hepatic regulation in dairy cows during transition and early lactation. *J. Dairy Sci.* 2014;97:5481–90.
33. Zarrin M, De Matteis L, Vernay MCMB, Wellnitz O, van Dorland HA, Bruckmaier RM. Long-term elevation of  $\beta$ -hydroxybutyrate in dairy cows through infusion: effects on feed intake, milk production, and metabolism. *J. Dairy Sci.* 2013;96:2960–72.
34. Kusuhara K, Yamamoto K, Okada K, Mizuno Y, Hara T. Association of IL12RB1 polymorphisms with susceptibility to and severity of tuberculosis in Japanese: a gene-based association analysis of 21 candidate genes. *Int. J. Immunogenet.* 2007;34:35–44.
35. Graber M, Kohler S, Kaufmann T, Doherr MG, Bruckmaier RM, van Dorland HA. A field study on characteristics and diversity of gene expression in the liver of dairy cows during the transition period. *J. Dairy Sci.* 2010;93:5200–15.
36. Browning BL, Browning SR. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* 2009;84:210–23.
37. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. *Nucleic Acids Res.* 2012;41:D48–55.
38. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. *PLoS Genet.* 2006;2:e190.
39. Zhang J-T. Approximate and Asymptotic Distributions of Chi-squared-type Mixtures With Applications. *J. Am. Stat. Assoc.* 2005;100:273–85.
40. Freytag S, Bickeböller H. Comparison of three summary statistics for ranking genes in genome-wide association studies. *Stat. Med.* 2013;n/a-n/a.

# **CHAPTER 3 WHOLE LIVER TRANSCRIPTOME ANALYSIS REVEALS NOVEL FACTORS INVOLVED IN THE METABOLIC ADAPTATION OF THE TRANSITION COW**

*N.-T. Ha<sup>1,2</sup>, C. Drögemüller<sup>3</sup>, F. Schmitz-Hsu<sup>4</sup>, R. M. Bruckmaier<sup>1</sup>, H. Simianer<sup>2</sup>, J. J. Gross<sup>1</sup>*

<sup>1</sup> Veterinary Physiology, Vetsuisse Faculty University of Bern, Switzerland

<sup>2</sup> Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University Goettingen, Germany

<sup>3</sup> Institute for Genetics, Vetsuisse Faculty University of Bern, Switzerland

<sup>4</sup> Swissgenetics, Zollikofen, Switzerland

## Abstract

**Background:** During their early lactation, dairy cows experience a severe metabolic load often resulting in the development of various diseases. The inevitable deficiency in nutrients and energy at the onset of lactation requires the adjustment of the hepatic metabolism to overcome metabolic stress. Numerous studies have been carried out to elucidate the complex process of adaptation, which, until now, is not fully understood. In this study, we conducted a whole liver transcriptome analysis for the transition cow to identify novel factors crucial for the metabolic adaptation.

**Results:** Liver samples were taken from six cows at three time-points: T1 = week 3 *ante-partum*; T2 = week 2 *post-partum*; T3 = week 3 *post-partum*. Using RNA-seq, we studied the transcriptomic profile of the transition cow before and after lactation. We performed a differential gene expression (DGE) and a combination of the gene-set enrichment analysis and perturbation analysis for biological pathways (KEGG database). Among the ~10,000 expressed genes, we discovered ~1,000 genes to be significantly differentially expressed ( $FDR < 5\%$ ), of which ~43% and ~16% are linked to lipids and oxidative stress, respectively, but only ~6% to the glucose metabolism (GO database). The combined pathway analysis further revealed 7 pathways to be significantly associated with the hepatic changes of the transition cow, including ‘adipocytokine signaling pathway’ and ‘steroid hormone biosynthesis’.

**Conclusion:** The DGE and pathway (biological gene set) analysis demonstrated that major hepatic changes from late pregnancy to early lactation are related to gluconeogenesis and fat mobilization (‘adipocytokine signaling pathway’). We further found indications for immunological changes (*GPX3*, ‘steroid hormone biosynthesis’ and the associated *CYP* and *UGT* transcripts) that may contribute to the impaired immune system of dairy cows during the transition period. The outcome of this study provides new insights into the un-

derstanding of the metabolic adaptation within the liver of the transition cow, which should be more closely investigated in future studies.

**Keywords:** RNA-seq, transition cow, early lactation, hepatic transcriptome, differential gene expression analysis

## Background

The transition period of a dairy cow, mostly defined as the time frame from 3 weeks *ante-partum* until 3 weeks *post-partum*, is characterized by abrupt changes in physiology, metabolism as well as nutrition of the animal [1,2]. Intensive fetal growth before parturition, morphological and endocrine changes related to mammary gland development, and the tremendous increase in energy and nutrient demand with onset of lactation are the most important challenges which the dairy cow has to face during this period. Hence, an optimal metabolic adaptation is required in order to avoid the development of metabolic and infectious diseases.

The liver is a key organ controlling and regulating metabolic adaptation. Therefore, a number of studies have been carried out to quantify the molecular adaptations during the transition period [3–10] in order to enhance the understanding of the complex biology of the transition cow. The main idea of these studies was to assess the transcriptomic profile of candidate genes known to be involved in the liver metabolism during transition using either mRNA abundances [6,8,9] or real-time PCR [7]. Whereas most of these studies focused mainly on specific metabolic processes, a more extensive study was conducted by Loor et al. [3] included the gene expression profiles of more than 6,300 genes using microarray technology. Even though these studies are valuable contributions to the understanding of the metabolic adaptation, they are still limited in the number of genes as well as on pre-selected genes, which may prevent detection of new regulators influencing the hepatic adaptation.

According to our knowledge, we are one of the first studies to use RNA-sequencing (RNA-seq) to compare the whole liver transcriptome of dairy cows from late-pregnancy to early lactation. This study may, on the one hand, validate the results established by earlier studies and, on the other hand, reveal new possible candidate genes crucial for the hepatic adaptation. A similar approach has been performed by McCabe et al. [11], who analyzed the whole liver transcriptome using RNA-seq to compare cows divergent in negative energy balance (NEB). However, in practice, the extent of NEB is not fully indicative for the fact whether a cow can successfully adapt to the new physiological state of lactation or not. Indeed, cows with an optimal adaptive performance would even overcome an extremely severe NEB without any occurrence of health disorders, whereas more vulnerable cows are more prone to fail even under less challenging situations [12–14].

## Results

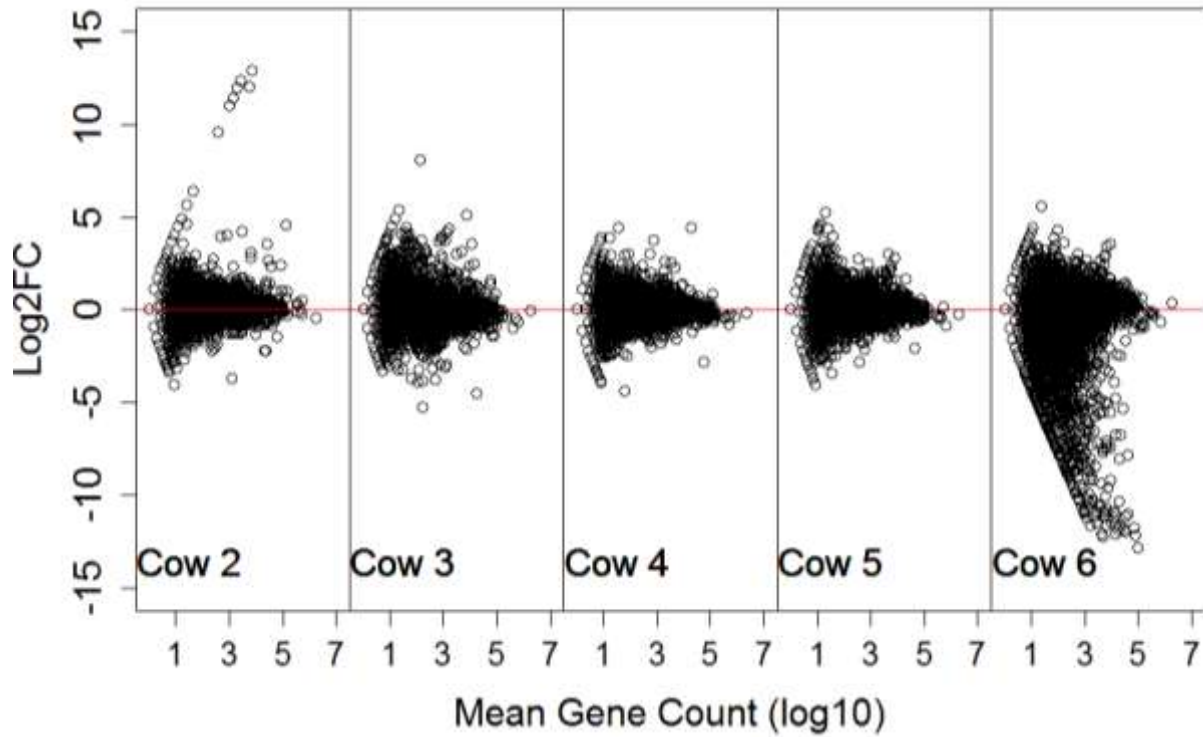
### *RNA-sequencing Data*

**Table 3.1:** Experimental design of the 17 RNA-sequencing libraries. Note that there is no sample of cow 1 at T1.

Cow-ID	Cow1	Cow2	Cow3	Cow4	Cow5	Cow6
3 weeks <i>ante-partum</i> (T1, $21 \pm 5$ days <i>ante-partum</i> )	-	L12	L13	L14	L15	L16
2 weeks <i>post-partum</i> (T2, $10 \pm 2$ days <i>post-partum</i> )	L21	L22	L23	L24	L25	L26
3 weeks <i>post-partum</i> (T3, $17 \pm 2$ days <i>post-partum</i> )	L31	L32	L33	L34	L35	L36

In order to investigate the hepatic changes during the transition period, we collected 17 liver samples of 6 dairy cows at 3 points of time (Table 3.1). Before filtering and analysis of the RNA-seq data, we performed a simple comparison between the cows to avoid any bias resulting from outliers in the samples. Indeed, when comparing the log2-fold-changes of the gene counts in T3 vs. T1 for the five cows, we discovered at least two cows with

gene expression patterns differing substantially from the rest (Figure 3.1). Note that the sample for cow 1 at T1 is missing and hence, a comparison is not possible for this cow.



**Figure 3.1:** Log2-foldchanges of the gene counts for all genes in T3 compared to T1. The x-axis depicts the logarithm of the mean gene counts.

On the one hand, we found seven genes to be highly up-regulated with fold-changes (FC) higher than 500 for cow 2. On the other hand, cow 6 showed significantly down-regulated genes with fold-changes up to 5000. The seven up-regulated genes in cow 2 were only found to be expressed in the sample L32. As for the gene expression analysis, we will only consider genes that were expressed in at least four of the 17 samples in order to avoid unreliable results. Hence, the sample L32 was still included in the following analyses.

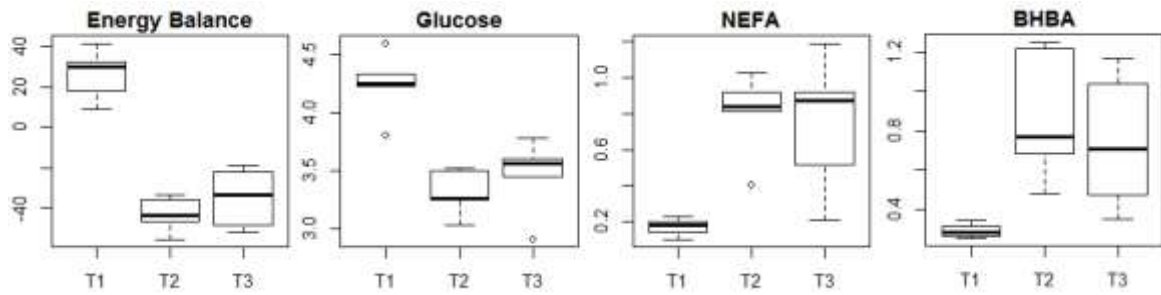
We further studied the gene expression patterns of cow 6 and found sample L16 to be an outlier. Supplementary Table 3.1 lists the interesting genes that were highly overexpressed in T1 for cow 6. We performed a simple Wilcoxon-rank-sum test to examine whether these genes are enriched in certain KEGG pathways [15] and discovered many significant path-

ways involved in several human diseases and myocardial disorders (supplementary Table 3.2), suggesting these findings to have a biological background. Suspecting that cow 6 may have undergone special conditions before calving, we excluded it from the following analyses. Indeed, cow 6 was bought from a private farm, and an earlier infection or health disorders before drying off were not reported. However, the present data might reflect health disturbances in anamnesis. After checking for outliers, we filtered the remaining 14 samples in order to keep only genes with more than 5 counts per one million reads (cpm) in at least 4 of the 14 samples, resulting in a set of 10,186 genes.

#### *Performance and Metabolic Status of Cows*

**Table 3.2:** Milk yield on test day, energy balance in the test week and metabolic status of the studied cows. The average day of sampling relative to parturition is -21 (+/- 4.6) for T1, 9.5 (+/-1.6) for T2 and 16.5 (+/- 1.6) for T3.

Time	Sample	Day of sampling (relative to parturition)	Milk yield (kg)	Energy balance (MJ NEL)	Glucose (mmol/L)	NEFA (mmol/L)	BHBA (mmol/L)
T1	L12	-16		8,89	3,81	0,14	0,31
	L13	-24		41,29	4,23	0,10	0,25
	L14	-25		31,95	4,6	0,23	0,34
	L15	-24		17,91	4,25	0,18	0,28
	L16	-16		29,83	4,34	0,2	0,26
T2	L21	+9	32,1	-41,27	3,52	0,92	0,48
	L22	+12	35,8	-45,20	3,28	0,81	0,81
	L23	+7	30,8	-55,21	3,25	1,03	0,68
	L24	+9	30,0	-33,39	3,50	0,40	0,72
	L25	+10	30,1	-35,76	3,03	0,83	1,22
	L26	+10	31,0	-46,10	3,25	0,85	1,25
T3	L31	+16	36,0	-47,85	3,57	0,52	0,47
	L32	+19	38,8	-51,78	3,79	0,21	0,35
	L33	+14	32,2	-43,08	3,61	1,19	0,57
	L34	+16	31,6	-22,08	3,45	0,86	0,84
	L35	+17	36,0	-18,79	2,91	0,88	1,04
	L36	+17	38,3	-23,43	3,56	0,92	1,17



**Figure 3.2:** Boxplots of energy balance (EB, MJ NEL/d), plasma NEFA (mmol/L), BHBA (mmol/L) and glucose for the six studied cows in T1, T2 and T3.

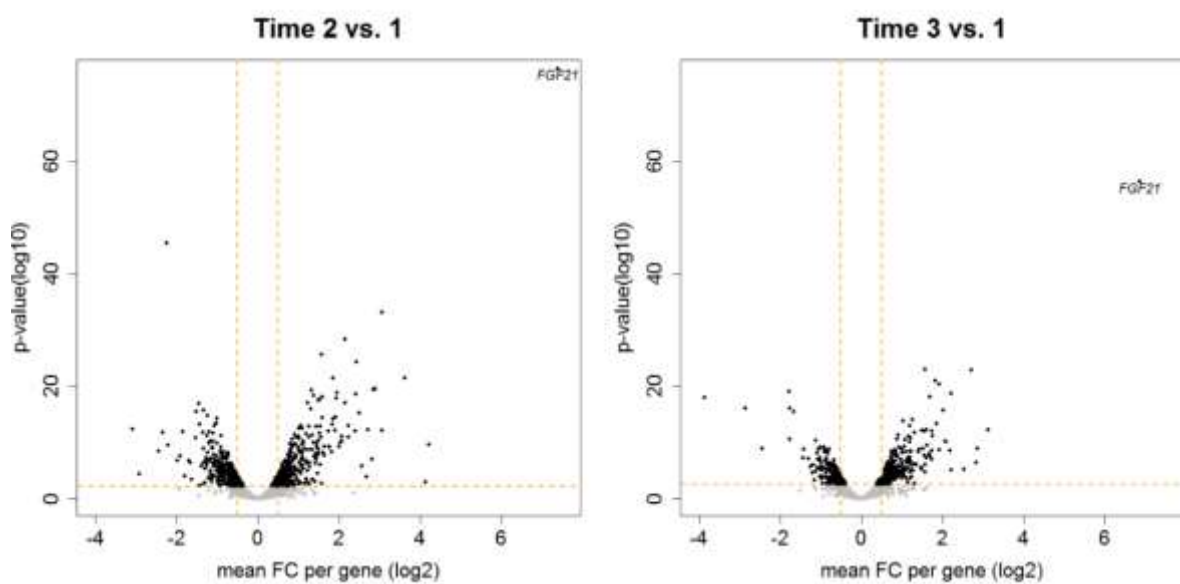
During late gestation at T1, energy balance (EB) of all cows was positive, but became negative during lactation (T2 and T3, Table 3.2). As expected T2 and T3 present similar values, while EB is slightly higher in T3. This effect, however, is not significant, while the changes from T1 to T2 and T1 to T3 are both significant with p-values of  $p=0.0007$  and  $p=0.00009$ , respectively. The changes in metabolic state of the six cows follow the typical pattern of a transition cow. With the onset of lactation, elevated plasma concentration of NEFA and BHBA are expected, coupled with a decrement of plasma glucose (Figure 3.2). Interestingly, the plasma NEFA and BHBA exhibit substantial between-cows variation in T2 and T3 than in T1 ( $\sigma_{T1,NEFA} = 0.05$ ,  $\sigma_{T2,NEFA} = 0.21$ ,  $\sigma_{T3,NEFA} = 0.34$ ) indicating the differences in their adaptive performance.

#### *Differential Gene Expression Analysis*

We conducted a differential gene expression analysis comparing T1 vs. T2 and T1 vs. T3. As expected, the two comparisons yielded very similar results, emphasizing the metabolic and physiologic similarity of the second and third week *post-partum* when cows are experiencing a distinct NEB (Figure 3.3). Nevertheless, Figure 3.3 shows that the hepatic changes in T2 are slightly higher than in T3 compared to T1, indicating a time-dependency of the metabolic changes on the transcriptomic level and being consistent with the observed metabolic status of the cows. All in all, the numbers of significantly differentially expressed genes are 967 (493 up-regulated) in T1 vs. T2 and 493 (318 up-regulated) in T1



vs. T3, respectively, at a false discovery rate (FDR) of 5% (compare supplementary Table 3.3 and 3.4), of which 397 genes were significant in both comparisons. The majority of the significant genes are related to known processes occurring after calving, e.g. the mobilization of lipids. This is reflected by the fact that about 43% of the 632 significant genes (in T1 vs. T2 or T1 vs. T3) with FC greater than 1.5 are linked to lipids; about 16% to oxidative activities, but only about 6% are linked to glucose metabolism (supplementary Table 3.5) according to the gene ontology (GO) database [16,17].



**Figure 3.3:** Volcano-plots of the results of the DGE analysis. Each dot in the plot represents a gene with its corresponding logFC on the x-axis and p-value (log10) on the y-axis. The horizontal line indicates the significance threshold (FDR<5%), whereas the vertical line segregates genes with logFC>1.5.

### *Pathway Analysis*

The gene-set enrichment analysis (GSEA) using the permutation-based weighted Kolmogorov-Smirnov test (WKS) [18] in combination with the perturbation analysis (PEA) [19] for KEGG pathways revealed 6 and 4 significant pathways (FDR < 0.05) in the comparisons T1 vs. T2 and T1 vs. T3, respectively. Table 3.3 lists the details of the WKS and PEA as well as the combined p-value from Fisher's method for all significant pathways. For a

detailed list of the results for all pathways, please refer to the supplementary Tables 3.6 and 3.7 for the comparisons T1 vs. T2 and T1 vs. T3, respectively. Note that the PEA is only applicable for signaling pathways with existent gene interaction information. Among the seven significant pathways, only one had gene network information for the perturbation analysis, namely the pathways ‘adipocytokine signaling pathway’.

**Table 3.3:** Results of the GSEA and perturbation analysis for the 7 significant pathways (FDR<5%) in T1 vs. T2 and T1 vs. T3 with their number of annotated genes and p-values. Note that a perturbation analysis is only applicable for pathways with present gene network information.

T1 vs. T2	Size	GSEA Score	GSEA p-value	Pert Factor	Pert p-value	Combined p-value
Steroid hormone biosynthesis	38	0,59	1,00E-04	0	NA	1,00E-04
Arachidonic acid metabolism	34	0,60	0,0002	0	NA	0,0002
Retinol metabolism	36	0,57	0,0004	0	NA	0,0004
Adipocytokine signaling pathway	46	0,54	0,0002	2,87	0,23	0,0005
Galactose metabolism	17	0,67	0,0008	0	NA	0,0008
Linoleic acid metabolism	16	0,64	0,0015	0	NA	0,0015
T1 vs. T3	Size	GSEA Score	GSEA p-value	Pert Factor	Pert p-value	Combined p-value
Steroid hormone biosynthesis	38	0,55	0,0004	0	NA	0,0004
Arachidonic acid metabolism	34	0,54	0,0009	0	NA	0,0009
Adipocytokine signaling pathway	46	0,50	0,0021	4,31	0,048	0,0010
Citrate cycle (TCA cycle)	28	0,56	0,0012	0	NA	0,0012

## Discussion

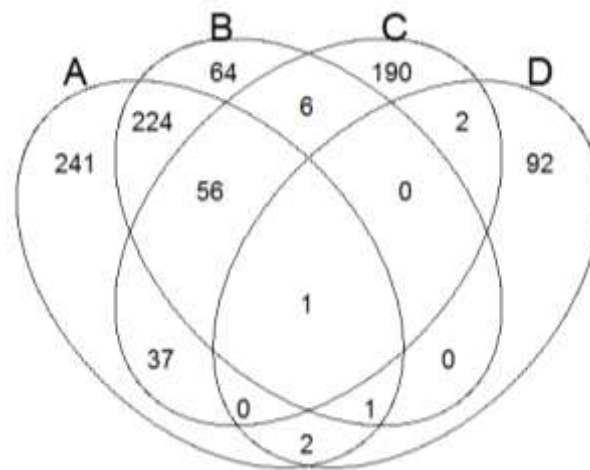
### *Statistical Analysis*

We performed a differential gene expression analysis using the R-package ‘edgeR’ followed by two pathway analyses, namely the GSEA by Subramanian et al. [18] and signaling pathway impact analysis (SPIA) by Tarca et al. [19]. The SPIA approach consists of two independent steps: (1) an over-representation analysis (ORA) to determine whether a certain pathway is enriched by significant genes and (2) a so-called perturbation analysis (PE) to capture the impact of gene expression changes to the considered pathway using its topology structure. The main idea, hereby, is that expression changes in certain genes that highly interact with other genes may have a higher probability to perturb the function of the pathway than down-stream genes. The final p-value for a pathway is then given by the combination of the ORA and perturbation analysis. However, the ORA method implemented by Tarca et al. [19] is Fisher’s exact test, which requires a declaration of an arbitrary significant threshold for the differentially expressed genes, which may exclude potentially informative genes, a main and known drawback of ORA approaches. To this end, Subramanian et al. [18] introduced the GSEA approach to overcome these limitations by ranking all genes according to their strength of association (e.g. p-value or FC). Therefore, in our study, we consider a combination of the two approaches in order to make use of their advantages (for further details, please refer to the section Material and Methods).

### *Comparison With Other Studies*

We further compared our results with those of Ha et al. [20], where a genome-wide association study was performed for the key metabolites glucose, NEFA, and BHBA during the transition period, and McCabe et al. [11] (Figure 3.4). We found 99 genes to be significant ( $\text{FDR} < 0.05$  and  $\text{FC} > 1.5$ ) in both our studies and in McCabe et al. [11]. Surprisingly, there are more than 700 significant genes that were only detected by McCabe et al. [11] or

the present study, of which about 500 genes are unique to our study. To figure out which types of genes are among these genes, we conducted a simple Fisher's Exact test to identify GO terms [17] enriched with the ~500 genes unique to our study. As a result, we find five significant GO terms at a FDR of 0.05 (supplementary Table 3.8). According to these results, we find the majority of these genes to be involved in either regulating processes supporting the breakdown of triglyceride as a result of fat mobilization in early lactation or biosynthetic processes for the formation of sterols, in particular cholesterol, and their metabolism. The increased activity of genes involved in hepatic cholesterol synthesis is in accordance with the observations by Schlegel et al. [8] and Kessler et al. [21] supporting their hypothesis that this adaptation aims to provide the mammary gland with cholesterol and triglycerides for milk production, which was until then only known for other species, such as rats [8,22]. The uniqueness of these genes as compared to McCabe et al. [11] further indicates that these processes might be common to all lactating cows regardless of their NEB.



**Figure 3.4:** Venn-diagram of the number of significant genes comparing the results of our analysis T1 vs. T2 (A) and T1 vs. T3 (B) with that of McCabe et al. (C) and Ha et al. (D). As for Ha et al. the number refers to the number of significant genes at a FDR of 5%, whereas for McCabe et al. and the results of this study, the numbers refer to the number of significant genes at a FDR of 5% and fold change greater than 1.5.

The gene *WFDC2* (*Bos Taurus WAP four-disulfide core domain 2*), known to be involved in the negative regulation of the peptidase activity according to GO [17], was significant in all three studies. However, until now, the functionality of this gene with respect to the metabolic adaptation of the transition cow is not understood. *WAP* (*Whey acidic protein*) is a major milk protein of many species, e.g. rats and wallabies, but is known to be absent in milk of cows, sheep and humans [23]. Nevertheless, the *WFDC2* protein was recently detected to have an antibacterial activity in the mammary gland of the tammar-wallaby during lactation [24]. Therefore, it cannot be ruled out that this gene may have a function in the hepatic adaptation, since it has been identified to be significantly differentially expressed (down-regulated, > 1.5-fold) in our study and significantly associated with the transition cow in two other independent studies [11,20].

In all three studies the pathway “steroid hormone biosynthesis” was found to be highly significant. As McCabe et al. [11] already mentioned, the liver usually is known for steroid hormone inactivation. In our study, 11 genes within this pathway were significantly differentially expressed with four genes (*LOC100296421*, *UGT2B10*, *HSD3B1*, *CYP11A1*) highly up-regulated and the rest (*LOC785540*, *CYP2C87*, *SULT1E1*, *CYP1A1*, *CYP7B1*, *HSD17B2*, *UGT2A3*) being down-regulated. Similar to McCabe et al. [11], we detected an up-regulation of *CYP11A1*, a member of the cytochrome P450 (*CYP*) family and a key enzyme in the metabolism of cholesterol to pregnenolone [11,25]. However, in contrast to their study, we even find further up-regulated genes, namely *LOC100296421*, *UGT2B10*, *HSD3B1*. It is known that *HSD3B1* (*type I 3 $\beta$ -hydroxysteroid dehydrogenases*) is crucial for the biosynthesis of hormonal steroids, such that aldosterone, cortisol and testosterone [26], whereas *LOC100296421* (*UGT2B4*) and *UGT2B10* are members of the *UDP* glucuronosyltransferase (*UGT*) 2 family, which is known for the metabolism and elimination of androgens in humans [27]. Together with the *CYPs*, *UGTs* are responsible for the metabolism of most hepatically cleared drugs in humans [28]. Further, Congiu et al. [29] detected

a decrease in *UGT* mRNA expression which was associated with the degree of inflammation for three isoforms of the *UGT* family including the significant *UGT2B4* ( $FC > 1.5$ ). This highly suggests these genes to be involved in the increased susceptibility of dairy cows to infectious diseases during their transition period.

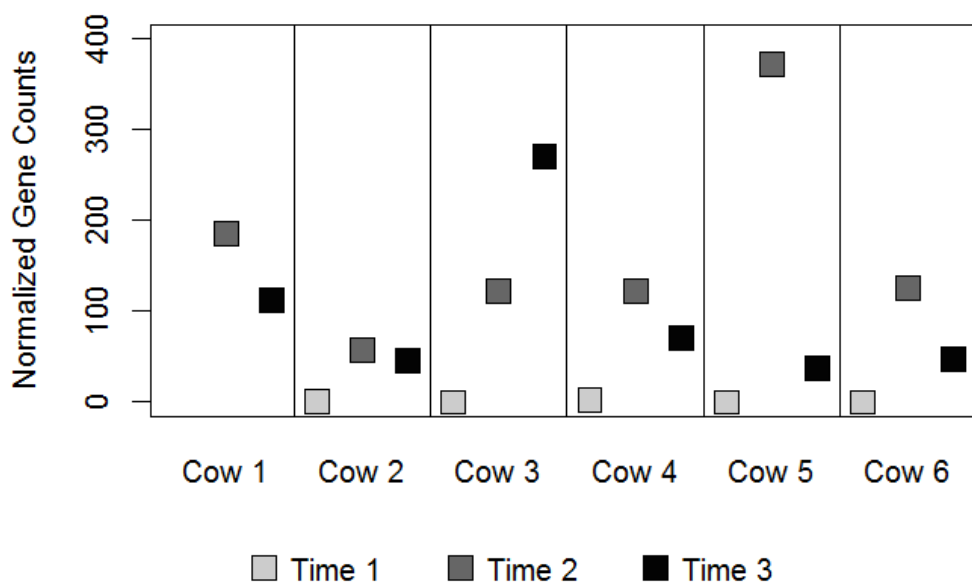
#### *Hepatokine FGF21*

The most striking gene is *FGF21* (fibroblast growth factor 21) with p-values of  $5.3 \times 10^{-77}$  and  $1.8 \times 10^{-52}$  in T1 vs. T2 and T1 vs. T3, respectively. Compared to the other significant genes having up to 12-fold-changes, this gene exhibits fold-changes greater than 167 in T1 vs. T2 and 111 in T1 vs. T3 (Figure 3.3). Indeed, this finding is consistent with the observations of recent studies [8,30–34], where *FGF21* gene expression is substantially altered during the transition period of dairy cows. Whereas Carriquiry et al. [30] hypothesized the functionality of this gene to be in the uncoupling of the ST-IGF axis, others found evidence for an inhibitory effect of *FGF21* on GH signaling [33] and its relevance for ketotic cows [31].

Further studies have suggested *FGF21* to be associated with the excessive use of lipid reserves [12,34] and to be relevant to many health-related aspects during transition and lactation in dairy cows, e.g. rate of gluconeogenesis,  $\beta$ -oxidation of free fatty acids, and development of ketosis [13,35]. In particular, Schoenberg et al. [34] detected that elevated *FGF21* expression is stimulated by the energy deficit during early lactation. Even though it was not directly mentioned in McCabe et al. [11], *FGF21* was also significantly different ( $FC > 2.49$   $p = 6.9 \times 10^{-09}$ ) between cows with a severe and mild NEB in their study. Moreover, in Carriquiry et al. [30] the mRNA expression of *FGF21* was found to be negatively correlated with EB and plasma glucose, but positively with plasma NEFA concentration. However, in our study, we failed to establish a significant correlation of the gene expres-

sion pattern (of T2 and T3, log-transformed) and the NEB (sample correlation  $-0.32$ ,  $p=0.317$ ), probably due to the very limited sample size.

Figure 3.5 illustrates the gene expression pattern of *FGF21* for the six dairy cows. In general, the graph indicates the absence of the *FGF21* gene expression in T1 (cpm  $< 1$ ) and a substantial increase in expression in T2, which subsequently drops drastically in T3. The extent of this expression pattern is highest for cow 5, but is completely modified for cow 3 showing a steady increase in gene expression from T1 to T3. This substantial between-cow variation of the hepatic *FGF21* expression was also found in Schoenberg et al. [34] and may be a reason for the lack of correlation for EB. Nevertheless, we discovered that the gene expression pattern of T2 and T3 (log-transformed) are significantly negatively correlated (sample correlation  $-0.74$ ,  $p=0.006$ ) with milk yield on the test day. Indeed, *FGF21* is known to increase glucose uptake in adipocytes in mice [36]. Together with our results, we suggest the up-regulation of this gene might be associated with the glucose demand of the mammary gland for milk production.



**Figure 3.5:** Gene expression pattern of the gene *FGF21* for the six cows at the three points of time.

### *Gluconeogenesis*

The main metabolic changes happening from late pregnancy to early lactation are related to the onset of lactation. Hence, we expected to observed major hepatic changes in the regulation of gluconeogenesis. However, the KEGG pathway ‘glycolysis / gluconeogenesis’ itself was not significantly affected after multiple testing correction ( $p = 0.087$  in T1 vs. T2 and  $p = 0.115$  in T1 vs. T3), which is also the case in McCabe et al. [11]. However, in contrast to their study, we effectively compare non-lactating vs. lactating cows and thus, differences were expected. Indeed, even though it is not significant, we discovered many genes within this pathway to be significantly up-regulated (*ALDOA*, *G6PC*, *FBP2*, *LDHB*, *LDHA*), whereas *PKLR*, a major rate limiting enzyme in gluconeogenesis [11] is significantly down-regulated. Furthermore, the up-regulation of *LDHB* and *LDHA* suggest a greater activity of the liver pyruvate dehydrogenase and the overexpression of G6PC in T2 and T3 as compared to T1 clearly shows the enhanced activity of the gluconeogenesis as a response to the high glucose demand for milk production.

This finding is further confirmed by a significant up-regulation of the precursor pyruvate carboxylase (PC, FC>2). However, as opposed to Graber et al. [6], we found no differences in the expression of propionyl-CoA carboxylase (*PCC*)  $\alpha$  and  $\beta$  and *PEPCK-m*, but solely a trend for a possible up-regulation of *PEPCK-c* (cytosolic *PCK1*,  $p = 0.029$ ) that controls the entry of amino acids and propionate into the tricarboxylic acid cycle (TCA cycle or citrate cycle). This observation is consistent with the significance of the TCA cycle with 6 significantly up-regulated genes (*SUCLA2*, *OGDH*, *IDH2*, *IDH1*, *IDH3A*, *PC*) suggesting a higher activity of this cycle as a result of the intense breakdown of fatty acids during the transition period. Interestingly, the observed trend in *PCK1*, but not in *PCK2* (*PEPCK-m*) is similar to the results of Agca et al. [37], but are completely opposed to that of Ostrowska et al.[7] and van Dorland et al. [14]. In chickens, Watford et al. [38] proposed a model, where *PCK1* is required for gluconeogenesis from amino acids, whereas



*PCK2* is responsible for gluconeogenesis from lactate. Following this model, the discrepancies of the different studies including ours, may be explained by the individual adaptations of the cows as a response to the availability of different resources (e.g. amino acids, propionate, lactate) for gluconeogenesis.

#### *Significant Pathways Indicate Immunological Functionality*

Among the seven significant pathways, we discovered the three pathways ‘arachidonic acid metabolism’, ‘retinol metabolism’, and ‘linoleic acid metabolism’ to be significantly down-regulated. Here, we note that all significantly differentially expressed genes within the ‘linoleic acid metabolism’ are also contained in the ‘arachidonic acid metabolism’. As for the retinol (vitamin A) metabolism, we identified 6 (*LOC785540*, *CYP2C87*, *AOX1*, *CYP1A1*, *CYP26A1*, *UGT2A3*) out of the 9 genes in this pathway to be down-regulated, whereas half of the significant genes in the ‘arachidonic acid metabolism’ were down-regulated.

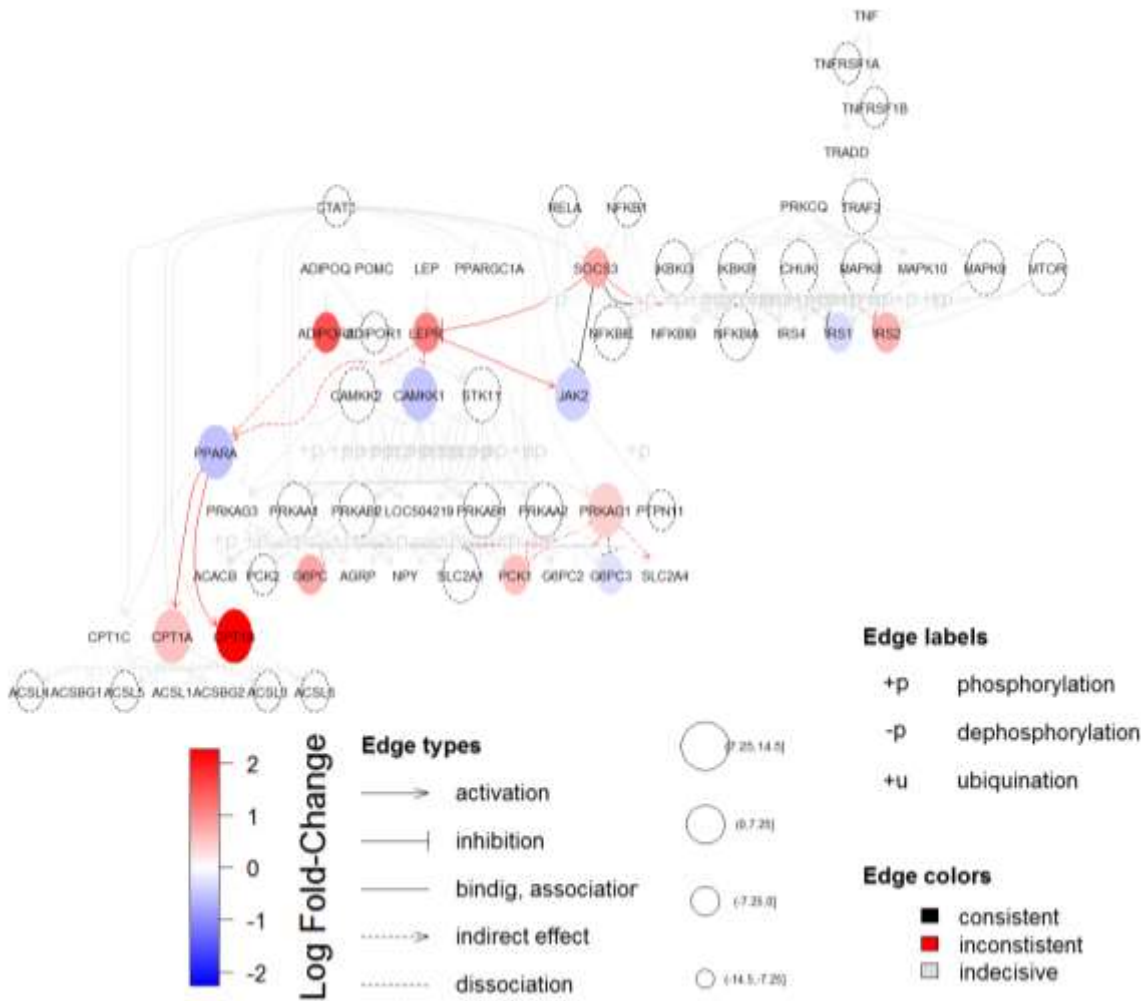
It was shown that low concentration of vitamin A and arachidonic acids may contribute to an impaired immune function [39]. Nevertheless, within the pathway ‘Arachidonic acid metabolism’ we found the gene *GPX3* to be significantly up-regulated ( $p=8.53^{-11}$ , FC > 12). As a member of the glutathione peroxidase enzyme family, it is an important defense mechanism against oxidative stress by protecting cells from oxidative damage [40]. Indeed, a number of studies have considered oxidative stress as one major factor for the success of metabolic adaptation [41–43]. In particular, Sordillo and Aitken [42] emphasized the need to enhance the antioxidant defense mechanisms of dairy cows during the transition period. In fact, our data exhibits a significant negative correlation between the *GPX3* expression pattern (log-transformed) and the EB (T2 and T3) with a sample correlation of -0.73 ( $p = 0.007$ ). This suggests that cows with a higher energy deficit are exposed to a higher rate of oxidative stress and thus exhibit higher level of plasma glutathione peroxi-

dase enzymes. Furthermore, we detected at least 16 additional genes that were significantly (FDR<5%) up-regulated (FC > 1.5) and were associated with anti-oxidative activities according to GO [17]. *GPX3*, the most up-regulated gene compared to other genes with anti-oxidative activities, is known to interact with selenium, an important dietary trace element closely linked to vitamin E metabolism and involved in numerous anti-oxidative processes.

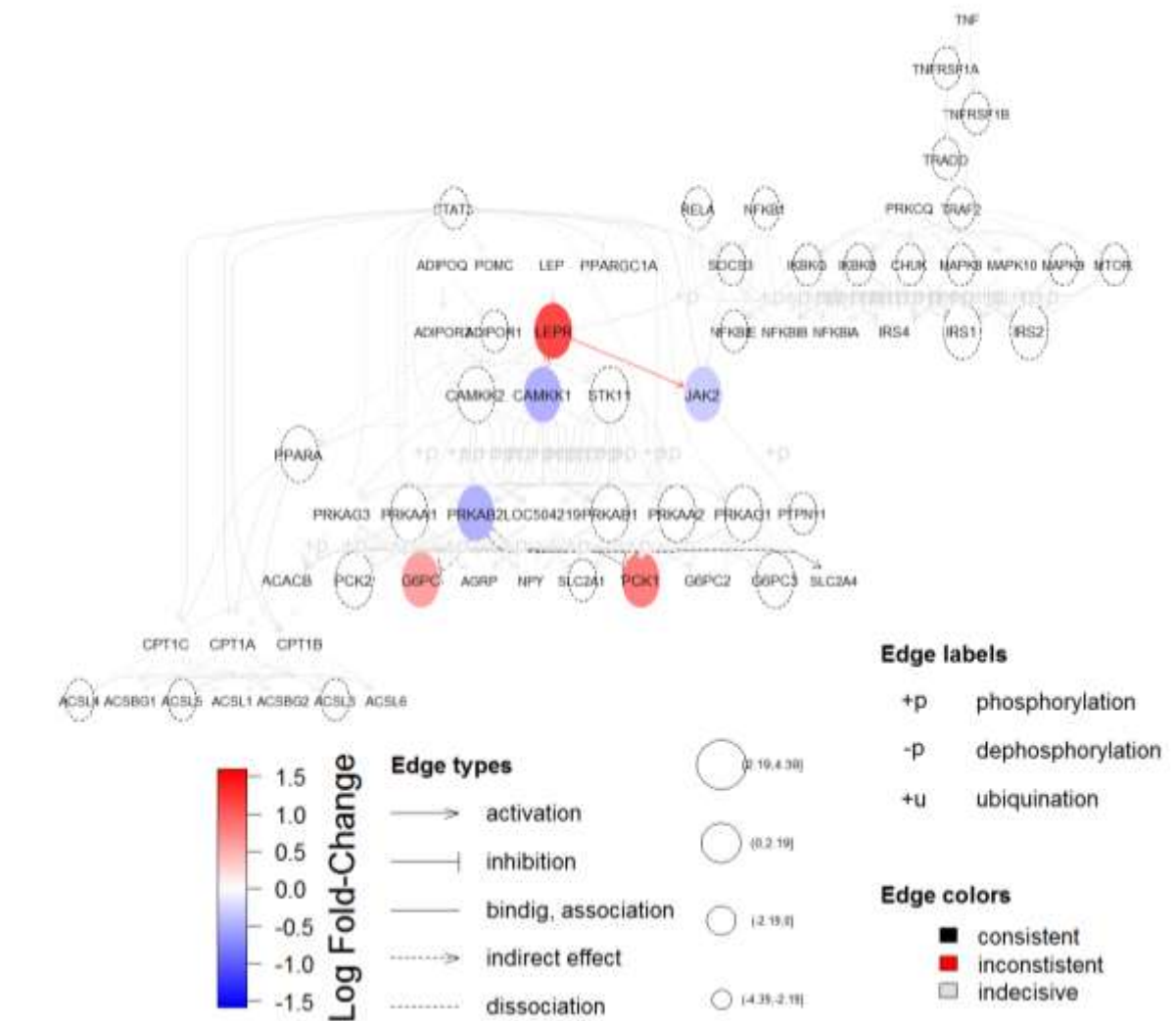
#### *Adipocytokine Signaling Pathway*

One of the most interesting findings is the ‘adipocytokine signaling pathway’. Among the 63 genes that were annotated to this pathway according to the KEGG database, we found 46 genes to be expressed (> 5 cpm) during the transition period, 14 of which are significantly differentially expressed, with nine genes being up-regulated after calving. Figures 3.6 and 3.7 show the results of the SPIA for this pathway for the comparisons T1 vs. T2 and T1 vs. T3, respectively. Red arrows indicate an inconsistency of the gene expression data to the KEGG network information. As illustrated by Figure 3.5, the ‘adipocytokine signaling pathway’ is highly perturbed in the transition period. Due to the overexpression of the two receptors for leptin (*LEPR*) and adiponectin (*ADIPOR2*), we expected to observe an activation of *PPARα*, which in turn, up-regulates *CPT1A* and *B* in order to enable the oxidation of adipose tissue derived fatty acids. However, contrary to the natural flow of this pathway, we detected a significant down-regulation of the *PPARα* gene expression (Figure 3.5) in T2. This significance totally vanishes in T3 (Figure 3.6). *PPARα* is the primary subtype of the peroxisome proliferator-activated receptors that are key factors in the regulation of fatty acid metabolism and gene transcription [30]. It is activated by a broad spectrum of fatty acids including linoleic acid, which indicates a possible relationship between the down-regulation of the pathway ‘linoleic acid metabolism’, as discussed earlier, and the down-regulation of *PPARα*. The reports in literature on this issue are so far inconsistent: in contrast to the present study, Loor et al. [3] observed an increase in *PPARα* expression until day 49 after calving. In Carriquiri et al. [30] the increase in expression was

only observed after 136 DIM, whereas Khan et al. [33] find that *PPARα* expression decreases around parturition. Nevertheless, we still observed an increment of *PPARα* targets, among others the genes *FGF21*, *CPT1A*, *CPT1B* as discussed above.



**Figure 3.6:** Map of the pathway “adipocytokines signaling pathway” created based on the perturbation analysis[19] and the R-package TOPASeq[63] for the comparison T1 vs. T2. Genes with dashed boundaries are expressed but not significantly differentially expressed. Black arrows indicate that the gene-to-gene interaction is consistent with our expression data, whereas red arrows indicate inconsistency.



**Figure 3.7:** Map of the pathway “adipocytokines signaling pathway” created based on the perturbation analysis[19] and the R-package TOPASeq[63] for the comparison T1 vs. T3. Genes with dashed boundaries are expressed but not significantly differentially expressed. Black arrows indicate that the gene-to-gene interaction is consistent with our expression data, whereas red arrows indicate inconsistency.

The ‘adipocytokine signaling pathway’ was shown to play a key role in lipid metabolism and feed efficiency [44]. Considering the role of leptin and insulin (state of insulin resistance after parturition, respectively) changes of the ‘adipocytokine signaling pathway’ enable the supply of energy and nutrients during the stage of a negative energy balance. The early lactation period is characterized by a decreased rate of lipogenesis and an in-

creased rate of lipolysis [45]. Adipose tissue (AT) is the major site for lipolysis, during which NEFA are released to support energy supply. However, AT is not only a source for fatty acids, but also an important site for synthesis and secretion of adipokines, a collective term for various AT derived hormones, proteins, and other bioactive molecules [46,47]. Predominately during periods of NEB, gene expression of leptin was shown to be reduced [48]. Besides leptin, adiponectin is one of the major circulating adipokines, and plays an important role in the regulation of glucose and fatty acid metabolism [47,49]. Furthermore, by acting on insulin sensitivity [50], adiponectin is supposed to be a key regulator of nutrient partitioning during early lactation [47]. Singh et al. [47] observed an inverse association between plasma adiponectin and NEFA concentrations during a NEB in dairy cows at different lactation stages and assumed inhibitory effects of adiponectin on lipolysis and stimulatory effects on hepatic fatty acid oxidation underlying their findings, as shown by earlier studies [50,51]. Whereas Singh et al. [47] speculated on a potential feed intake stimulatory effect of adiponectin by the AMP-activated protein kinase as shown thitherto only for mice [52], our present pathway analysis for the ‘adipocytokine signaling pathway’ clearly indicates a causal association between adipokine and AMPK expression in transition dairy cows. The earlier observations by Singh et al. [31] of independent regulatory functions of adiponectin and leptin are confirmed by the changes of single components within the overall pathway by our present investigations.

## Conclusions

The substantial hepatic changes (~10% of all expressed genes) from late-pregnancy to early lactation emphasize the complexity of the metabolic adaptation in dairy cows. Nevertheless, we were able to identify further genes and pathways that, to our knowledge, have not been linked to the transition cow on a transcriptomic level. In particular, we found indications for immunological changes (*GPX3*, ‘steroid hormone biosynthesis’ and the associated

*CYP* and *UGT* transcripts) that may contribute to the impaired immune system of dairy cows during the transition period. The discovered factors are a valuable contribution to the understanding of the metabolic adaptation and hence, should be more closely investigated in future studies.

## Material and Methods

### *Characterization of Cows and Liver Sampling*

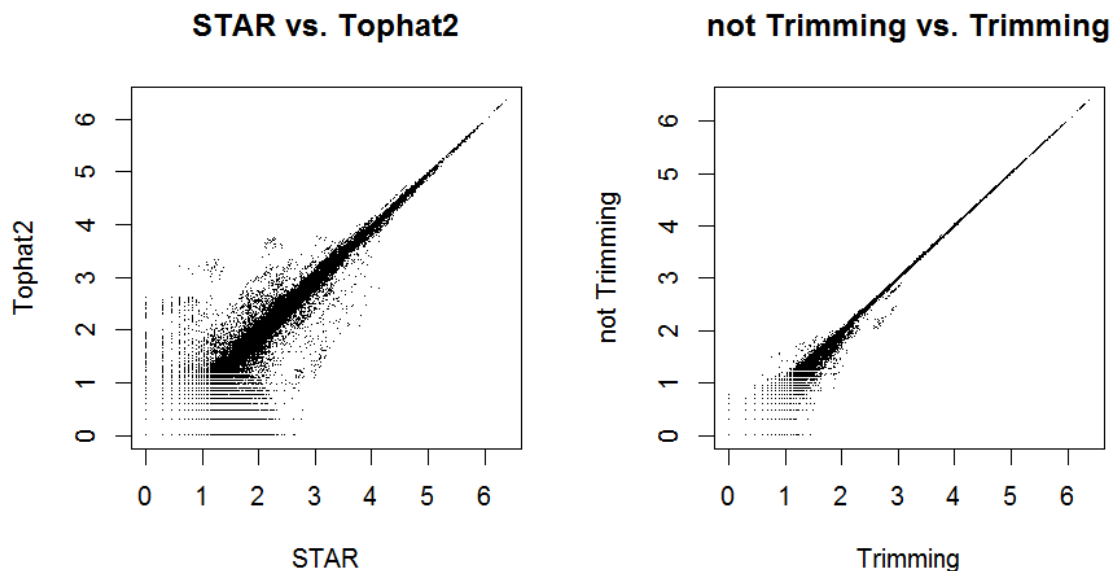
For our analysis, we collected liver and blood samples from 6 dairy cows (Red Holstein) at three different stages: 3 weeks before expected calving (T1,  $22 \pm 4$  days *ante-partum*), 2 weeks ( $10 \pm 2$  days *post-partum*) and 3 weeks ( $17 \pm 2$  days *post-partum*) after calving (T2 and T3, respectively). Cows were kept at an experimental station with individual daily recording of feed intake and milk yield. Liver sampling was performed by blind percutaneous needle biopsy under local anesthesia as described by van Dorland et al. [14]. Liver samples were directly put in RNeasy lysis buffer (Qiagen, Applied Biosystems Business, Austin TX, USA) and stored at  $-80^{\circ}\text{C}$  until extraction. Following milk yield, dry matter intake, and development of body weight, EB of the animals was calculated on a weekly basis. Blood samples concomitantly obtained prior to liver sampling were analyzed for concentrations of glucose, non-esterified fatty acids (NEFA), and  $\beta$ -hydroxybutyrate (BHBA) to characterize the metabolic changes occurring during the transition period. Apart from the liver samples, the 6 cows were also phenotyped for various traits and metabolic characteristics. Table 3.2 shows the extract of the phenotype data that were used in this study, including the milk yield and energy balance.

### *RNA Isolation and Sequencing*

RNA was isolated from liver samples using RNeasy Mini kit (Qiagen). The quality and quantity of the isolated RNA was measured with an Agilent 2100 Bioanalyzer (Agilent Technologies) and Qubit 2.0 Fluorometer (Life Technologies). The RNA integrity number

(RIN) of all samples was between 8.1 and 9.0. Approximately 800 ng of high quality RNA was used for strand-specific paired-end RNA library preparation (TruSeq stranded mRNA sample preparation guide Part #15031047 Rev.D, Illumina). Total mRNA libraries were multiplexed (pool of 4 uL of each sample after dilution to 2 nM, adding ca. 1% PhiX control) and sequenced each in three lanes (in the same run) on the Illumina HiSeq2000 platform using 2 x 100 bp paired-end sequencing cycles. The Illumina BCL output files with base calls and qualities were converted into FASTQ file format and demultiplexed with the CASAVA (v1.8.2) software (Illumina).

### *RNA Read Alignment and Counting*



**Figure 3.8:** Comparison of the different mapping methods. Each dot represents the number of read pairs ( $\log_{10}$ ) assigned to the same gene and sample for different mapping methods, i.e. STAR vs. Tophat (not trimmed) and trimming vs. not trimming (STAR).

**Table 3.4:** The mean rate of mapped and counted read pairs across all samples for the different methods. The numbers are percentages of the average total number of read pairs N.

N = 23,015,133	STAR		Tophat
	not trimmed	trimmed	not trimmed
rate of mapped pairs	0.94	0.915	0.796
rate of counted pairs	0.739	0.718	0.635

In order to avoid bias in the subsequent analyses, we conducted a comprehensive quality control using FastQC [53] (version 0.11.2) for the forward and backward reads of the three lanes for all 17 samples. In general, none of the read sequences failed the overall quality control; nevertheless, as a matter of course, there are sequences containing bases with low qualities ( $<10$  Sanger Quality Score). In order to assess whether these outliers do have an impact on our analyses, we trimmed the bad bases for all samples using the program Trimmomatic [54] (version 0.33). After quality control, we combined the different lanes for each sample before mapping them to the reference genome *Bos Taurus* (UMD3.1, release 78). For comparison purposes, we performed the alignment on the original as well as trimmed samples using two different programs, STAR [55] (Version 2.4) and Tophat2 [56] (version 2.0.13), based on a gene annotation (*Bos Taurus*, UMD3.1, release 78) downloaded from Ensembl [57] to improve the mapping performance. STAR and Tophat were run with their default parameters, since they are claimed to be optimized for mammalian species (compare the corresponding manuals). Finally, we used featureCounts [58] (version 1.4.6-p19) to count the number of reads overlapping with a specific gene. In order to avoid an inflation of false positives, we only counted read pairs that are uniquely mapped in a concordance manner and can be unambiguously assigned to an annotated gene. Table 3.4 depicts the average rate of mapped and counted read pairs for the different methods, whereas a comparison between trimming vs. not trimming, and STAR vs. Tophat2 are illustrated in Figure 3.8. For a complete list of the mapped and counted pairs for all samples, please refer to Table 3.5. All in all, we detected that the effect of trimming is negligible, while STAR has a higher mapping rate than Tophat2. Therefore, in the present paper, all results are based on the mapping program STAR for untrimmed data. For the differential gene expression analysis, we removed genes with less than 5 cpm (counts per 1 million reads) in at least 4 samples to avoid unreliable results. Finally, a matrix with counts of 10,186 genes for 17 samples was on hand.



**Table 3.5:** The total number and rate (bracket) of uniquely mapped and counted read pairs for each 17 samples for the final data (mapped with STAR, not trimmed).

Sample	# read pairs	# mapped pairs	# counted pairs
L12	27,339,731	25,479,412 (0.932)	19,756,739 (0.723)
L13	26,648,264	25,315,159 (0.950)	20,025,837 (0.751)
L14	23,475,475	22,134,917 (0.943)	17,679,582 (0.753)
L15	20,833,859	19,748,273 (0.948)	15,576,240 (0.748)
L16	24,250,944	22,956,153 (0.947)	18,570,700 (0.766)
L21	23,342,617	21,833,370 (0.935)	17,127,338 (0.734)
L22	21,724,762	20,360,111 (0.937)	15,794,314 (0.727)
L23	19,546,096	17,777,418 (0.91)	13,938,476 (0.713)
L24	23,416,586	21,869,282 (0.934)	17,201,584 (0.735)
L25	20,321,033	19,235,832 (0.947)	15,060,712 (0.741)
L26	24,075,637	22,750,704 (0.945)	17,925,743 (0.745)
L31	20,395,477	19,255,991 (0.944)	15,006,595 (0.736)
L32	22,092,573	20,506,446 (0.928)	16,006,660 (0.725)
L33	25,332,449	23,820,903 (0.940)	18,489,979 (0.730)
L34	21,855,757	20,618,961 (0.943)	16,312,584 (0.746)
L35	23,651,113	22,374,014 (0.946)	17,609,780 (0.745)
L36	22,954,884	21,690,343 (0.945)	17,031,698 (0.742)
Average	23,015,133	21,631,017 (0.940)	17,006,739 (0.739)

### *Differential Gene Expression and Pathway Analysis*

All analyses were conducted in R [59]. For the identification of genes differentially expressed before and after lactation, we employed the R-package ‘edgeR’ [60,61] using a paired design to account for the baseline differences between the cows. For each gene  $g$ , we used a likelihood-ratio test (“glmFit”) based on the generalized linear model

$$\log(Y) = X\beta + \log N,$$

where  $Y = (y_{Ti,1}, \dots, y_{Ti,n}, y_{Tj,1}, \dots, y_{Tj,n})^t$  and  $N = (N_{Ti,1}, \dots, N_{Ti,n}, N_{Tj,1}, \dots, N_{Tj,n})^t$ , are the normalized gene counts and library sizes for each cow in  $Ti$  and  $Tj$ ;  $X$  is the design matrix with the columns *Intercept* and the binary designed variables  $Cow_1, \dots, Cow_n$ , *Time* and  $\beta$  is the vector of regression coefficients. Using the negative binomial distribution for the gene counts, we test the null hypothesis that the coefficient  $\beta_{Time}$  is equal to zero comparing  $T1$  vs  $T2$  and  $T1$  vs  $T3$ .

We used the obtained p-values to perform a gene-set enrichment analysis (GSEA) based on the weighted Kolmogorov-Smirnov Test (WKS) suggested by Subramanian et al. [18] to detect pathways enriched with significant genes. To this end, we created a gene to pathway annotation based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [15]. As for the WKS, we made use of the weights  $-\log_{10}(p_j)$ , where  $p_j$  is the p-value of the of gene  $g_j$  obtained from the DGE analysis, and 10,000 permutations. For further details, please refer to Ha et al. [20] or Subramanian et al. [18].

We further conducted a signaling pathway impact analysis (SPIA) by Tarca et al. [19]. The SPIA approach consists of two independent steps: (1) an over-representation analysis (ORA) to determine whether a certain pathway is enriched by significant genes and (2) a perturbation analysis (PEA) to capture the impact of gene expression changes to the considered pathway using its topology structure. In our study, we consider a combination of the WKS and the PEA, i.e. the second part of the SPIA. For each pathway, we obtain its final p-value by combining the p-values of the WKS and PEA using Fisher's method [62].

The PEA as well as the visualization of the pathway data were conducted using the R-package "ToPASeq" [63], which is adapted for RNA-seq data. Since PEA is a network-based approach, it is only applicable for signaling pathways with gene network information (inhibition/activation). For those pathways, where a perturbation analysis is not applicable, we only consider the p-values obtain from the WKS approach.

Both the DGE and pathway analysis require the testing of multiple null hypothesis and thus, need to be properly accounted for. Here, we employed the false discovery rate (FDR) approach by Benjamini and Hochberg [64]. As for the pathway analysis, we reduced the multiple testing problems by filtering out non-relevant pathways, which are the KEGG pathways for human diseases.

### **Ethics Approval and Consent to Participate**

The experimental procedures followed the Swiss Federal Law on Animal Protection and were approved by the Committee of Animal Experiments of the Canton Fribourg, Switzerland.

### **Availability of Data and Material**

The dataset supporting the conclusions of this article is available in the NCBI SRA repository, SRP066548 (<http://www.ncbi.nlm.nih.gov/bioproject/PRJNA303131/>).

### **Competing Interests**

The authors declare no conflict of interest.

### **Funding**

This study was supported by a grant of the Swiss Commission for Technology and Innovation CTI (project no. 13948.2 PFLS-LS) and Swissgenetics, Zollikofen, Switzerland.

### **Authors' Contributions**

NH carried out the data preparation, statistical analyses and drafted the manuscript. JIG and RB designed the study and performed the experiments. CD participated in the RNA sequencing. FS and HS participated in the design of the study and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

## Acknowledgement

We thank Michèle Ackermann and Muriel Fragnière for expert technical assistance, and the Next Generation Sequencing Platform of the University of Bern for performing the sequencing experiments.

## Supplementary Tables

**Supplementary Table 3.1:** Gene information for the 343 genes with  $\log_{2}FC > 5$  in sample L16 (T1 and cow 6).  
<https://figshare.com/s/7168726c01d048afebbe>  
 doi:10.6084/m9.figshare.3369799

**Supplementary Table 3.2:** Pathways that were significantly enriched with up-regulated genes in T1 for cow 6.  
<https://figshare.com/s/8106f335d6f807d28347>  
 doi:10.6084/m9.figshare.3369802

**Supplementary Table 3.3:** List of all significant genes ( $FDR < 5\%$ ) from the DGE analysis comparing T1 vs. T2.  
<https://figshare.com/s/2d213d5f81a0906d2696>  
 doi:10.6084/m9.figshare.3369808

**Supplementary Table 3.4:** List of all significant genes ( $FDR < 5\%$ ) from the DGE analysis comparing T1 vs. T3.  
<https://figshare.com/s/b7628b467212d3daa6a0>  
 doi:10.6084/m9.figshare.3369811

**Supplementary Table 3.5:** Genes to gene ontology (GO) annotation of all significant genes ( $FDR < 5\%$ ) with  $FC > 1.5$  from the DGE analysis comparing T1 vs. T2 and T1 vs. T3.  
<https://figshare.com/s/55793bdea1f748d81a60>  
 doi:10.6084/m9.figshare.3369814

**Supplementary Table 3.6:** Results of the WKS and PEA for all pathways comparing T1 vs. T2. <https://figshare.com/s/2e8f8635579042e9b087>  
 doi:10.6084/m9.figshare.3369817

**Supplementary Table 3.7:** Results of the WKS and PEA for all pathways comparing T1 vs. T3. (<https://figshare.com/s/ef1405c63ba9158abf69>)  
doi:10.6084/m9.figshare.3369820

**Supplementary Table 3.8:** Results of the Fisher's Exact test for the ~500 genes unique to our study. The five GO terms [17] of the GO domain 'biological processes' are significant at a FDR of 0.05. (<https://figshare.com/s/93fcbb22b7bcff2f2fe0>)  
doi:10.6084/m9.figshare.3369823

## References

1. Gross J, van Dorland HA, Bruckmaier RM, Schwarz FJ. Performance and metabolic profile of dairy cows during a lactational and deliberately induced negative energy balance with subsequent realimentation. *J. Dairy Sci.* 2011;94:1820–30.
2. Gross J, van Dorland HA, Schwarz FJ, Bruckmaier RM. Endocrine changes and liver mRNA abundance of somatotrophic axis and insulin system constituents during negative energy balance at different stages of lactation in dairy cows. *J. Dairy Sci.* 2011;94:3484–94.
3. Loor JJ, Dann HM, Everts RE, Oliveira R, Green CA, Guretzky NAJ, et al. Temporal gene expression profiling of liver from periparturient dairy cows reveals complex adaptive mechanisms in hepatic function. *Physiol. Genomics.* 2005;23:217–26.
4. Reynolds CK, Aikman PC, Lupoli B, Humphries DJ, Beever DE. Splanchnic metabolism of dairy cows during the transition from late gestation through early lactation. *J. Dairy Sci.* 2003;86:1201–17.
5. Drackley JK, Dann HM, Douglas GN, Guretzky NAJ, Litherland NB, Underwood JP, et al. Physiological and pathological adaptations in dairy cows that may increase susceptibility to periparturient diseases and disorders. *Growth.* 2005;7.
6. Graber M, Kohler S, Kaufmann T, Doherr MG, Bruckmaier RM, van Dorland HA. A field study on characteristics and diversity of gene expression in the liver of dairy cows during the transition period. *J. Dairy Sci.* 2010;93:5200–15.
7. Ostrowska M, Górka BŻ, Słoniewski K, Kowalski LZ. Expression of PC, PCK1, PCK2, LDHB, FBP1 and G6PC genes in the liver of cows in the transition from pregnancy to lactation. *Anim. Sci. Pap. Rep.* 2013;31:281–90.
8. Schlegel G, Ringseis R, Keller J, Schwarz FJ, Eder K. Changes in the expression of hepatic genes involved in cholesterol homeostasis in dairy cows in the transition period and at different stages of lactation. *J. Dairy Sci.* 2012;95:3826–36.

9. Greenfield RB, Cecava MJ, Donkin SS. Changes in mRNA expression for gluconeogenic enzymes in liver of dairy cattle during the transition to lactation. *J. Dairy Sci.* 2000;83:1228–36.
10. Drackley JK, Overton TR, Douglas GN. Adaptations of Glucose and Long-Chain Fatty Acid Metabolism in Liver of Dairy Cows during the Periparturient Period. *J. Dairy Sci.* 2001;84, Supplement:E100–12.
11. McCabe M, Waters S, Morris D, Kenny D, Lynn D, Creevey C. RNA-seq analysis of differential gene expression in liver from lactating dairy cows divergent in negative energy balance. *BMC Genomics.* 2012;13:193.
12. Kessel S, Stroehl M, Meyer HHD, Hiss S, Sauerwein H, Schwarz FJ, et al. Individual variability in physiological adaptation to metabolic stress during early lactation in dairy cows kept under equal conditions. *J. Anim. Sci.* 2008;86:2903–12.
13. Gross JJ, R M Bruckmaier. Repeatability of metabolic responses to a nutrient deficiency in early and mid-lactation and its implication to robustness of dairy cows. *J. Dairy Sci.* 2015;
14. van Dorland HA, Richter S, Morel I, Doherr MG, Castro N, Bruckmaier RM. Variation in hepatic regulation of metabolism during the dry period and in early lactation in dairy cows. *J. Dairy Sci.* 2009;92:1924–40.
15. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012;40:D109–14.
16. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* 2000;25:25–9.
17. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015;43:D1049–56.
18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102:15545–50.
19. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J, et al. A novel signaling pathway impact analysis. *Bioinformatics.* 2009;25:75–82.
20. Ha N-T, Gross JJ, van Dorland A, Tetens J, Thaller G, Schlather M, et al. Gene-based mapping and pathway analysis of metabolic traits in dairy cows. *PloS One.* 2015;10:e0122325.
21. Kessler EC, Gross JJ, Bruckmaier RM, Albrecht C. Cholesterol metabolism, transport, and hepatic regulation in dairy cows during transition and early lactation. *J. Dairy Sci.* 2014;97:5481–90.

22. Smith JL, Lear SR, Forte TM, Ko W, Massimi M, Erickson SK. Effect of pregnancy and lactation on lipoprotein and cholesterol metabolism in the rat. *J. Lipid Res.* 1998;39:2237–49.
23. Sharp JA, Lefèvre C, Nicholas KR. Molecular evolution of monotreme and marsupial whey acidic protein genes. *Evol. Dev.* 2007;9:378–392.
24. Watt AP, Sharp JA, Lefevre C, Nicholas KR. WFDC2 is differentially expressed in the mammary gland of the tammar wallaby and provides immune protection to the mammary gland and the developing pouch young. *Dev. Comp. Immunol.* 2012;36:584–590.
25. Guryev O, Carvalho RA, Usanov S, Gilep A, Estabrook RW. A pathway for the metabolism of vitamin D3: Unique hydroxylated metabolites formed during catalysis with cytochrome P450scc (CYP11A1). *Proc. Natl. Acad. Sci.* 2003;100:14754–9.
26. Rosmond R, Chagnon M, Bouchard C, Björntorp P. Polymorphism in exon 4 of the human 3 $\beta$ -hydroxysteroid dehydrogenase type I gene (HSD3B1) and blood pressure. *Biochem. Biophys. Res. Commun.* 2002;293:629–32.
27. Chouinard S, Yueh M-F, Tukey RH, Giton F, Fiet J, Pelletier G, et al. Inactivation by UDP-glucuronosyltransferase enzymes: the end of androgen signaling. *J. Steroid Biochem. Mol. Biol.* 2008;109:247–53.
28. Williams JA, Hyland R, Jones BC, Smith DA, Hurst S, Goosen TC, et al. Drug-drug interactions for UDP-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (AUC<sub>i</sub>/AUC) ratios. *Drug Metab. Dispos. Biol. Fate Chem.* 2004;32:1201–8.
29. Congiu M, Mashford ML, Slavin JL, Desmond PV. UDP Glucuronosyltransferase mRNA Levels in Human Liver Disease. *Drug Metab. Dispos.* 2002;30:129–34.
30. Carriquiry M, Weber WJ, Fahrenkrug SC, Crooker BA. Hepatic gene expression in multiparous Holstein cows treated with bovine somatotropin and fed n-3 fatty acids in early lactation1. *J. Dairy Sci.* 2009;92:4889–900.
31. Akbar H, Batistel F, Drackley JK, Loor JJ. Alterations in Hepatic FGF21, Co-Regulated Genes, and Upstream Metabolic Genes in Response to Nutrition, Ketosis and Inflammation in Peripartal Holstein Cows. *PLoS ONE.* 2015;10:e0139963.
32. Akbar H, Schmitt E, Ballou MA, Corrêa MN, Depeters EJ, Loor JJ. Dietary Lipid During Late-Pregnancy and Early-Lactation to Manipulate Metabolic and Inflammatory Gene Network Expression in Dairy Cattle Liver with a Focus on PPARs. *Gene Regul. Syst. Biol.* 2013;7:103–23.
33. Khan MJ, Jacometo CB, Graugnard DE, Corrêa MN, Schmitt E, Cardoso F, et al. Overfeeding Dairy Cattle During Late-Pregnancy Alters Hepatic PPAR $\alpha$ -Regulated Pathways Including Hepatokines: Impact on Metabolism and Peripheral Insulin Sensitivity. *Gene Regul. Syst. Biol.* 2014;8:97–111.

34. Schoenberg KM, Giesy SL, Harvatine KJ, Waldron MR, Cheng C, Kharitononkov A, et al. Plasma FGF21 Is Elevated by the Intense Lipid Mobilization of Lactation. *Endocrinology*. 2011;152:4652–61.
35. Schlegel G, Ringseis R, Keller J, Schwarz FJ, Windisch W, Eder K. Expression of fibroblast growth factor 21 in the liver of dairy cows in the transition period and during lactation. *J. Anim. Physiol. Anim. Nutr.* 2013;97:820–9.
36. Ogawa Y, Kurosu H, Yamamoto M, Nandi A, Rosenblatt KP, Goetz R, et al. BetaKlotho is required for metabolic activity of fibroblast growth factor 21. *Proc. Natl. Acad. Sci. U. S. A.* 2007;104:7432–7.
37. Agca C, Greenfield RB, Hartwell JR, Donkin SS. Cloning and characterization of bovine cytosolic and mitochondrial PEPCK during transition to lactation. *Physiol. Genomics*. 2002;11:53–63.
38. Watford M, Hod Y, Chiao YB, Utter MF, Hanson RW. The unique role of the kidney in gluconeogenesis in the chicken. The significance of a cytosolic form of phosphoenolpyruvate carboxykinase. *J. Biol. Chem.* 1981;256:10023–7.
39. LeBlanc SJ, Herdt TH, Seymour WM, Duffield TF, Leslie KE. Peripartum serum vitamin E, retinol, and beta-carotene in dairy cattle and their associations with disease. *J. Dairy Sci.* 2004;87:609–19.
40. Mills GC. Hemoglobin Catabolism I. Glutathione Peroxidase, an Erythrocyte Enzyme Which Protects Hemoglobin from Oxidative Breakdown. *J. Biol. Chem.* 1957;229:189–97.
41. Bernabucci U, Ronchi B, Lacetera N, Nardone A. Influence of Body Condition Score on Relationships Between Metabolic Status and Oxidative Stress in Periparturient Dairy Cows\*. *J. Dairy Sci.* 2005;88:2017–26.
42. Sordillo LM, Aitken SL. Impact of oxidative stress on the health and immune function of dairy cattle. *Vet. Immunol. Immunopathol.* 2009;128:104–9.
43. Sharma N, Singh NK, Singh OP, Pandey V, Verma PK. Oxidative stress and antioxidant status during transition period in dairy cows. *Asian-Aust J Anim Sci.* 2011;24:479–84.
44. Xi YM, Yang Z, Wu F, Han ZY, Wang GL. Gene expression profiling of hormonal regulation related to the residual feed intake of Holstein cattle. *Biochem. Biophys. Res. Commun.* 2015;465:19–25.
45. McNamara JP, Hillers JK. Adaptations in lipid metabolism of bovine adipose tissue in lactogenesis and lactation. *J. Lipid Res.* 1986;27:150–7.
46. Kershaw EE, Flier JS. Adipose tissue as an endocrine organ. *J. Clin. Endocrinol. Metab.* 2004;89:2548–56.



47. Singh SP, Häussler S, Gross JJ, Schwarz FJ, Bruckmaier RM, Sauerwein H. Short communication: circulating and milk adiponectin change differently during energy deficiency at different stages of lactation in dairy cows. *J. Dairy Sci.* 2014;97:1535–42.
48. Block SS, Butler WR, Ehrhardt RA, Bell AW, Van Amburgh ME, Boisclair YR. Decreased concentration of plasma leptin in periparturient dairy cows is caused by negative energy balance. *J. Endocrinol.* 2001;171:339–48.
49. Yamauchi T, Kamon J, Minokoshi Y, Ito Y, Waki H, Uchida S, et al. Adiponectin stimulates glucose utilization and fatty-acid oxidation by activating AMP-activated protein kinase. *Nat. Med.* 2002;8:1288–95.
50. Yamauchi T, Kamon J, Waki H, Terauchi Y, Kubota N, Hara K, et al. The fat-derived hormone adiponectin reverses insulin resistance associated with both lipoatrophy and obesity. *Nat. Med.* 2001;7:941–6.
51. Qiao L, Kinney B, Schaack J, Shao J. Adiponectin inhibits lipolysis in mouse adipocytes. *Diabetes.* 2011;60:1519–27.
52. Kubota N, Yano W, Kubota T, Yamauchi T, Itoh S, Kumagai H, et al. Adiponectin stimulates AMP-activated protein kinase in the hypothalamus and increases food intake. *Cell Metab.* 2007;6:55–68.
53. Andrews, Simon. FastQC: a quality control tool for high throughput sequence data. 2010; Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
54. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics.* 2014;btu170.
55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics.* 2013;29:15–21.
56. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 2013;14:R36.
57. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. *Nucleic Acids Res.* 2012;41:D48–55.
58. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30:923–30.
59. R Core Team. R: A Language and Environment for Statistical Computing. 2014; R Foundation for Statistical Computing, Vienna, Austria. Available from: <http://www.R-project.org/>

60. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* 2010;26:139–40.
61. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res.* 2012;40:4288–97.
62. Mosteller F, Fisher RA. Questions and Answers. *Am. Stat.* 1948;2:30–1.
63. Ihnatova I, Budinska E. ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-seq data. *BMC Bioinformatics* [Internet]. 2015 [cited 2016 Jan 25];16. Available from: <http://www.biomedcentral.com/1471-2105/16/350>
64. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* 1995;57:289–300.

## **CHAPTER 4 CHARACTERIZING METABOLIC ROBUSTNESS OF DAIRY COWS DURING EARLY LACTATION USING REACTION-NORM MODELS**

*N.-T. Ha<sup>1,2</sup>, A. R. Sharifi<sup>2</sup>, M. Schlather<sup>3</sup>, U. Schnyder<sup>4</sup>, J. J. Gross<sup>1</sup>, F. Schmitz-Hsu<sup>5</sup>, R. M. Bruckmaier<sup>1</sup>, H. Simianer<sup>2</sup>*

<sup>1</sup> Veterinary Physiology, Vetsuisse Faculty University of Bern, Switzerland

<sup>2</sup> Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August-University Goettingen, Germany

<sup>3</sup> Chair of Mathematical Statistics, University of Mannheim, Germany

<sup>4</sup> Qualitas AG, Chamerstrasse 56, Ch-6300 Zug, Switzerland

<sup>5</sup> Senior Geneticist, Swissgenetics, Meielenfeldweg 12, 3052 Zollikofen, Switzerland

## **Abstract**

Dairy cows often suffer a severe metabolic stress in the transition phase, caused by the discrepancy of a high energy demand for rapidly increasing milk production and limited feed intake. We hypothesize that some cows are genetically less well suited to cope with this metabolic stress than others leading to adverse follow-up effects on longevity. To test this, we designed a reaction norm model in which the functional lifetime is linked to the metabolic challenge at the beginning of the first lactation. As challenge variables, we used either the sum of the milk yield or the accumulated fat/protein ratio of the first three test days, pre-adjusted for herd-test day variance. We defined a random regression sire model, in which a random regression term was modeled for each sire to assess the genetic component of the reaction to the challenge. The model was fitted to data of ~5,000 Brown Swiss bulls and ~580'000 daughters with suitable observations available ( $\geq 10$  daughters per bull). In order to validate our proposed model and to assess the reliability of the estimated (co)variance components, we conducted an extensive bootstrap approach. We found the sire variance for the slope of the random regression to be significantly different from zero for both challenge variables, suggesting a genetic component for the ability to cope with metabolic stress. The results of the study show that the ability to cope with metabolic stress in the transition phase clearly has a genetic component and can be used to breed metabolically robust dairy cows.

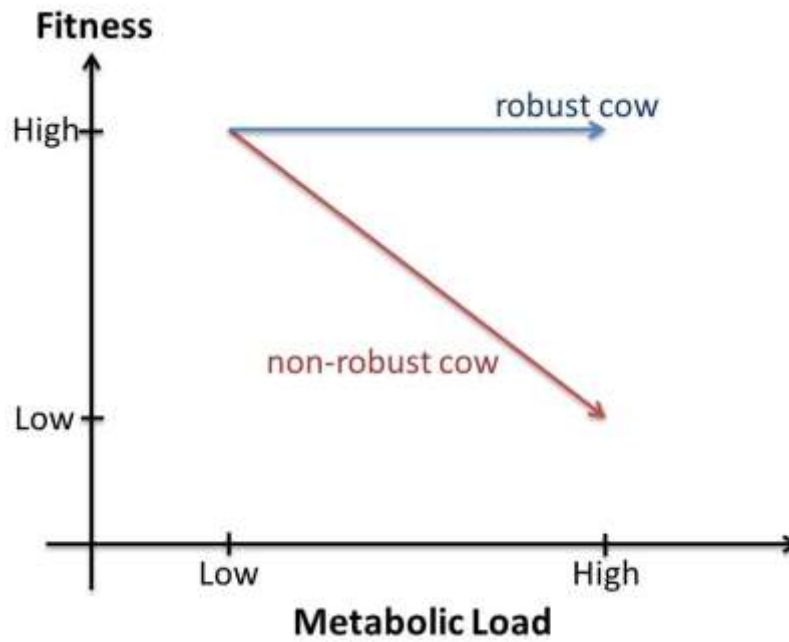
## **Introduction**

The phase of negative energy balance (NEB) which dairy cows inevitably undergo during early lactation has posed a major challenge for dairy science in the last decades. The substantial increase of energy demand due to the onset of lactation coupled with an insufficient feed intake makes dairy cows susceptible to major production-related diseases and infections in the early lactation phase [1–3]. An optimization of metabolic adaptation of

dairy cows would help to overcome this phase [4]. Despite the growing interest and number of studies aiming to unveil the complex mechanisms of this adaptation driven by the liver [5–13], practical management implications to effectively reduce the number of disorder occurrences during early lactation remain limited.

Even under the same conditions and similar production levels, the variability of how individual dairy cows deal with this situation is remarkably large [14,15]. It ranges from a successful metabolic adaptation without any clinical health problems to the development of production-related diseases, such as fatty liver and ketosis, coupled with a weakening of the immune system. This has led to the hypothesis that the metabolic adaptability may have a genetic basis [9,10,13,15,16]. If it exists, identifying this genetically determined metabolic robustness of dairy cows would allow us to breed for robust dairy cows that are better suited to cope with the excessive energy demand during the transition period. This goal, however, has not been reached so far, not least because there is no general definition of what metabolic robustness is and how it can be measured on a phenotypic scale.

In this study, we suggest defining the metabolic robustness of a dairy cow based on her reaction to a given metabolic load during early lactation. We further assume that metabolically robust cows are better able to cope with metabolic stress during early lactation. This ability is reflected by a generally increased fitness of the cow, which in the long run will result in a longer functional lifetime. Figure 4.1 presents a schematic representation of metabolic robustness and illustrates the difference between a robust and non-robust dairy cow. Robust dairy cows are expected to generally have a high fitness regardless of the extent of their metabolic load, whereas for non-robust dairy cows, the extent of metabolic stress has a negative effect on their fitness.



**Figure 4.1:** Schematic representation of the metabolic robustness.

In animal breeding, these kinds of models are referred to as reaction norm models and have been successfully applied for the analysis of genotype-by-environment interaction effects [17–19]. In this context, the metabolic load that a cow has to face during early lactation is termed the challenge variable and her fitness is called the response variable. The metabolic robustness is then defined as the slope of the fitted linear regression line of the response variable on the challenge variable. In order to fit such a model, repeated observations of the challenge and response variable would be required for each genotype, i.e. for each dairy cow.

However, in practice, each dairy cow is subjected to a single metabolic load in a given lactation and thus, fitting a regression line for each dairy cow is not possible. Hence, we suggest the use of a reaction norm sire model, in which a regression line is estimated for each bull using the data of his daughters exposed to different challenges and thus exhibiting different reactions. For our analysis, we assume that the fitness of a certain cow is reflected by her functional lifetime, whereas the challenge is represented by the energy deficit in the

early part of her first lactation, which is quantified by either the cumulated milk yield or the fat to protein ratio.

The aims of the present study are (1) to estimate breeding values for the metabolic robustness (slopes) using a reaction norm sire model and (2) to validate the proposed model by means of an extensive bootstrap approach, in which both the random effects and residuals are resampled.

## **Material and Methods**

### *Raw Data and Data Filtering*

In the present study, around 38 million test day records of ~1.9 million dairy cows of the breed Brown Swiss were available for the traits milk yield, fat and protein content. These test day records were pre-adjusted for herd-test day variance as well as length of gestation, which was done in the routine genetic evaluation system ([http://www.interbull.org/web/static/mace\\_evaluations\\_archive/eval/pr-dec10.html](http://www.interbull.org/web/static/mace_evaluations_archive/eval/pr-dec10.html)). Among these cows, around 1.4 million cows also had records for the trait “functional lifetime” obtained from the routine animal genetic evaluation ([http://www.interbull.org/web/static/mace\\_evaluations\\_archive/longevity/long-apr13.html](http://www.interbull.org/web/static/mace_evaluations_archive/longevity/long-apr13.html)) as pseudo-records and their corresponding weights from SurvivalKit V3.0 [20], based on dead and living animals born between 1976 and 2014. In general, the pseudo-record should reflect the longevity of an animal corrected for all non-genetic effects and the corresponding weight the reliability of its calculation [21]. In the following, the term functional lifetime is defined as the logarithmized pseudo-record. For our analysis, we filtered these data by only considering (1) cows that were born before 2011 and have died in order to have a reliable own record for longevity, (2) animals with pseudo-records having weights greater than 0.0005 in order to remove unreliable pseudo-records, and (3) animals with at least three test day records within the first 120 days in milk (DIM) of the first lactation to be able to calculate the challenge variable.

### *Challenge and Response Variables*

For each dairy cow, we defined functional lifetime as the response variable and used the sum of the first three test-day records in the first lactation either for milk yield or for the fat to protein ratio as challenge variable. Since the test-day records of the different cows may be on different DIMs, we performed a linear regression of the challenge on the sum of their first three DIMs and corrected for this effect if the slope was significant ( $p < 0.05$ ). To avoid unreliable estimates due to rare and extreme environments, we excluded extreme values (3% on both tails of the distribution) of the challenge variables and standardized them to have mean zero and standard deviation of one. Only bulls with at least 10 daughters with records were considered for analysis, resulting in a data set of 4,983 bulls and ~580,000 daughters with records. The number of daughters per bull ranges between 10 and 11,470 with a median of 30. In the following, these 4,983 bulls are referred to as the reference bulls.

### *Statistical Analyses*

We assessed the metabolic robustness of dairy cows using the following reaction-norm sire model

$$y_{ij} = \mu + \beta x_{ij} + s_i + b_i x_{ij} + \epsilon_{ij}, \quad (1)$$

where  $y_{ij}$  and  $x_{ij}$  are the response and challenge variables of the  $j$ -th daughter of sire  $i$ , respectively;  $\mu$  and  $\beta$  are the fixed coefficients for intercept and slope;  $s_i$  and  $b_i$  are the random intercept and slope for sire  $i$  with

$$\begin{pmatrix} s \\ b \end{pmatrix} \sim \left( 0, A \otimes \begin{pmatrix} \sigma_s^2 & \sigma_{sb} \\ \sigma_{sb} & \sigma_b^2 \end{pmatrix} \right)$$

and  $\epsilon_{ij} \sim (0, \frac{\sigma_e^2}{w_{ij}})$  is the scaled residual term as suggested by Ducrocq et al. [22] and Joaquim et al. [21], where  $w_{ij}$  are the weights of the pseudo-record  $y_{ij}$  as obtained from



SurvivalKit V3.0 [20] .  $A$  is the additive-genetic relationship matrix calculated based on the corresponding pedigree, constructed to include the 4,983 reference bulls and their ancestors up to the 10<sup>th</sup> generation and resulting in 30,842 animals for which breeding values were estimated. For comparison purposes, we further fitted a simple sire model based on the equation

$$y_{ij} = \mu + s_i^* + \epsilon_{ij}, \quad (2)$$

with the random errors  $\epsilon_{ij} \sim (0, \frac{\sigma_e^2}{w_{ij}})$  and random sire effects  $s^* \sim (0, \sigma_{s^*}^2 A)$  as to obtain an estimate for the heritability of the response variable when not accounting for challenge variables. All models were fitted using ASReml 3.0 [23].

In general, heritability is defined as the ratio between genotypic and phenotypic variation. As for model (2) without the genetic component of the challenge variables, the total genetic variance is given by four times the sire variance  $\hat{\sigma}_{s^*}^2$  and thus, heritability is defined by

$$h^2 = \frac{4\hat{\sigma}_{s^*}^2}{\hat{\sigma}_{s^*}^2 + \hat{\sigma}_e^2}.$$

In the reaction norm sire model (1), where the phenotype is modeled to be dependent on the challenge variable, heritability is also a function of the challenge variable  $x$  and is given by [24]

$$h^2(x) = \frac{4\hat{\sigma}_G^2(x)}{\hat{\sigma}_G^2(x) + \hat{\sigma}_e^2}$$

where  $\hat{\sigma}_G^2(x) = \hat{\sigma}_s^2 + \hat{\sigma}_b^2 x^2 + 2\hat{\sigma}_{sb}x$ .

In order to test whether the variance components  $\sigma_s^2$ ,  $\sigma_b^2$  and  $\sigma_{sb}$  are different from zero, we used the likelihood ratio test. To this end, four hierarchically nested sub-models of model (1) were fitted:

sub-model 1:  $\sigma_s^2 = 0, \sigma_b^2 = 0, \sigma_{sb} = 0$

sub-model 2:  $\sigma_s^2 > 0, \sigma_b^2 = 0, \sigma_{sb} = 0$

sub-model 3:  $\sigma_s^2 > 0, \sigma_b^2 > 0, \sigma_{sb} = 0$

sub-model 4:  $\sigma_s^2 > 0, \sigma_b^2 > 0, \sigma_{sb} \neq 0$ .

The test statistic for the null hypotheses  $H_0^{(1)}: \sigma_s^2 = 0$ ,  $H_0^{(2)}: \sigma_b^2 = 0$  and  $H_0^{(3)}: \sigma_{sb} = 0$  are given by  $T_1$ ,  $T_2$  and  $T_3$ , respectively, where  $T_i = 2(\log L_i - \log L_{i+1})$  and  $L_i$  is the likelihood of the fitted sub-model  $i$  obtained from ASReml 3.0 [23]. While the null distribution of  $T_3$  is a  $\chi_1^2$ -distribution, we used the  $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$  distribution for  $T_1$  and  $T_2$  to account for the fact that the corresponding null hypotheses lie on the boundary of the respective parameter space [25].

We further assessed the accuracy of the estimated variance components by means of a bootstrap method, while keeping the structure of our dataset as well as fixing the number of daughters per bull in the reference. To this end, we applied a method based on the bootstrapping of both random effects and residuals, which has been proven to have a superior performance for mixed models as compared to other, non-parametric approaches [26,27]. However, since the random effects  $s_i$  and  $b_i$  are correlated with a known covariance structure, which is the relationship matrix  $A$ , we had to account for this by adapting the bootstrap procedure as follows [26]:

- (1) Fit the reaction norm sire model to the original data to obtain the random effects  $\hat{s}_i$  and  $\hat{b}_i$  and residuals  $\hat{\epsilon}_{ij} = y_{ij} - \hat{\mu} - \hat{\beta}x_{ij} - \hat{s}_i - \hat{b}_i x_{ij}$  with their corresponding weights  $x_{ij}$
- (2) Sampling random effects:

- a. Transform the random effects  $s' = A^{-1/2}\hat{s}$ ,  $b' = A^{-1/2}\hat{b}$ , such that the set of  $\{(s'_i, b'_i)\} = \{(s'_1, b'_1), (s'_2, b'_2), \dots, (s'_m, b'_m)\}$  are independent tuples, where  $m$  is the total number of random effects or sires
- b. Draw a sample of tuples  $\{(s'^*_i, b'^*_i)\}$  of size  $m$  with replacement from  $\{(s'_i, b'_i)\}$
- c. Transform the random effects back using  $s^* = A^{1/2}s'^*$ ,  $b^* = A^{1/2}b'^*$

(3) Sampling residuals:

- a. Transform the residuals  $\epsilon'_{ij} = \frac{\hat{\epsilon}_{ij}}{w_{ij}}$ , such that the residuals  $\{\epsilon'_{ij}\}$  are independently and identically distributed
- b. Draw a sample  $\{\epsilon'^*_{ij}\}$  of size  $N$  with replacement from  $\{\epsilon'_{ij}\}$ , where  $N$  is the total number of data records
- c. Transform the residuals back using  $\epsilon^*_{ij} = w_{ij}\epsilon'^*_{ij}$

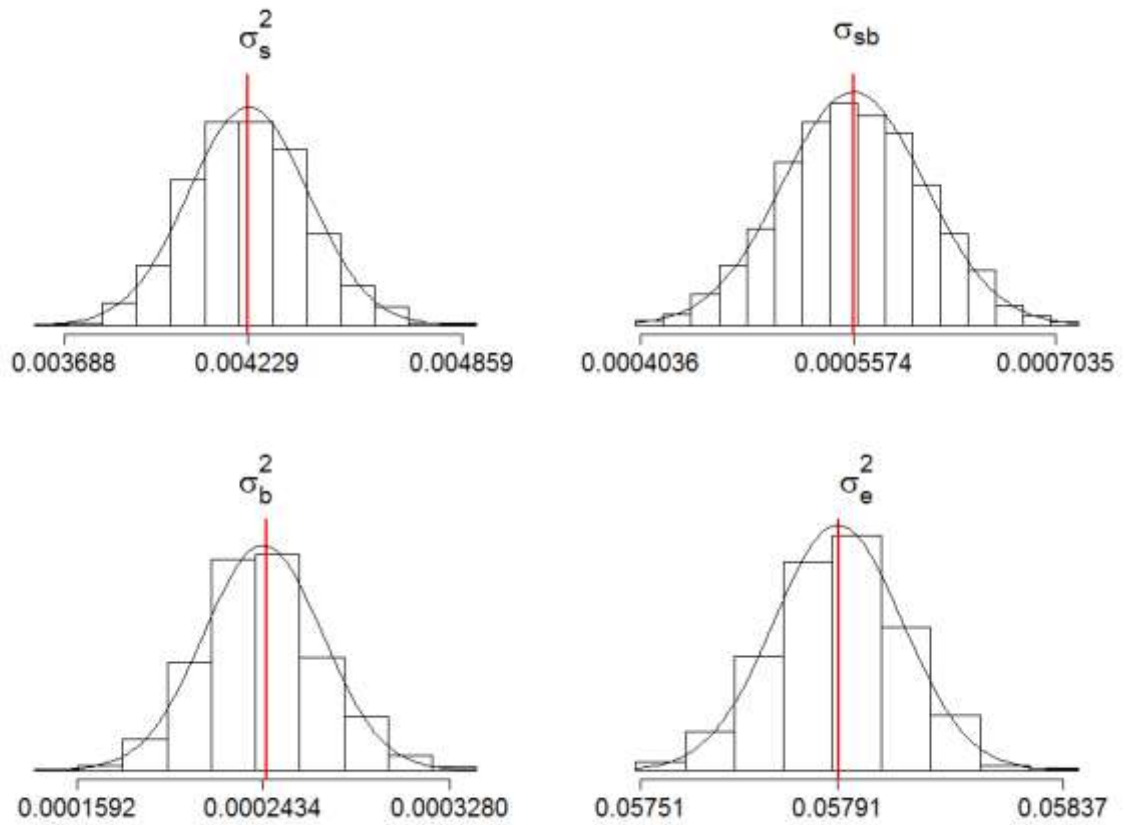
(4) Fit the reaction norm sire model to the new dataset  $y^*_{ij} = \hat{\mu} + \hat{\beta}x_{ij} + s^*_i + b^*_i x_{ij} + \epsilon^*_{ij}$  to obtain relevant variance components

Steps (1) to (4) were repeated 1,000 times, so that a sample of  $n = 1,000$  bootstrap estimators for all (co)variance components were available reflecting their empirical distribution. Based on this distribution, we calculated typical measures of accuracy such as variance, standard errors or confidence intervals. Here, we note that between step (2a) and (2b) as well as between step (3a) and (3b), we applied a further correction for both random effects and residuals to avoid a bias resulting from the difference between empirical and estimated variance components as shown by previous studies[26,28–31]. To this end, the random effects as well as residuals were first centered and then scaled with the ratio matrix of the estimated and empirical covariance matrices (for further details, please refer to [26]).

By sampling the random effects, we assumed that the random slopes are interchangeable between bulls. Therefore, this bootstrap procedure is inappropriate if we aim to assess the accuracy of the predicted slopes for each sire separately. In order to obtain confidence bands for the predicted slopes  $\hat{b}_i$ , we conducted another bootstrap procedure, where the covariance parameters  $\hat{\sigma}_s^2$ ,  $\hat{\sigma}_b^2$  and  $\hat{\sigma}_{sb}$  were fixed and given by the estimators obtained from the original data. For each reference bull  $i$  and his set of daughter records  $\{(y_{i1}, x_{i1}, w_{i1}), (y_{i2}, x_{i2}, w_{i2}), \dots, (y_{id_i}, x_{id_i}, w_{id_i})\}$ , we draw a bootstrap sample of size  $d_i$  with replacement. Using the new bootstrap sample of all reference bulls and the variance parameters  $\hat{\sigma}_s^2$ ,  $\hat{\sigma}_b^2$  and  $\hat{\sigma}_{sb}$ , a new prediction for the random effects including the random intercepts and slopes for all bulls were calculated. This procedure was repeated 1,000 times, such that in the end, a sample of  $n = 1,000$  estimators  $\hat{b}_i^{(1)}, \hat{b}_i^{(2)}, \dots, \hat{b}_i^{(n)}$  was available reflecting the empirical distribution of  $\hat{b}_i$ .

## Results

The results of the reaction norm model given by equation (1) for two different challenge variables based on milk yield and fat/protein ratio are presented in Table 4.1. As indicated by their standard errors obtained from ASReml 3.0 [23], all components are significantly different from zero except for the fixed parameter  $\beta$  for the challenge variable ‘milk’. This is further confirmed by the results of the log-likelihood ratio tests, where all genetic components were found to contribute significantly to a better fitting of the reaction norm model as compared to the model without the respective genetic component (all p-values  $< 10^{-8}$ ). The empirical distributions of the four variance components  $\hat{\sigma}_s^2$ ,  $\hat{\sigma}_b^2$ ,  $\hat{\sigma}_{sb}$  and  $\hat{\sigma}_e^2$  as obtained from the bootstrap approach are presented in Figure 4.2. The standard errors calculated based on these distributions are similar to those obtained directly from ASReml 3.0 (Table 4.1) with minor deviations. More importantly, according to Figure 4.2, all variance estimators are unbiased.

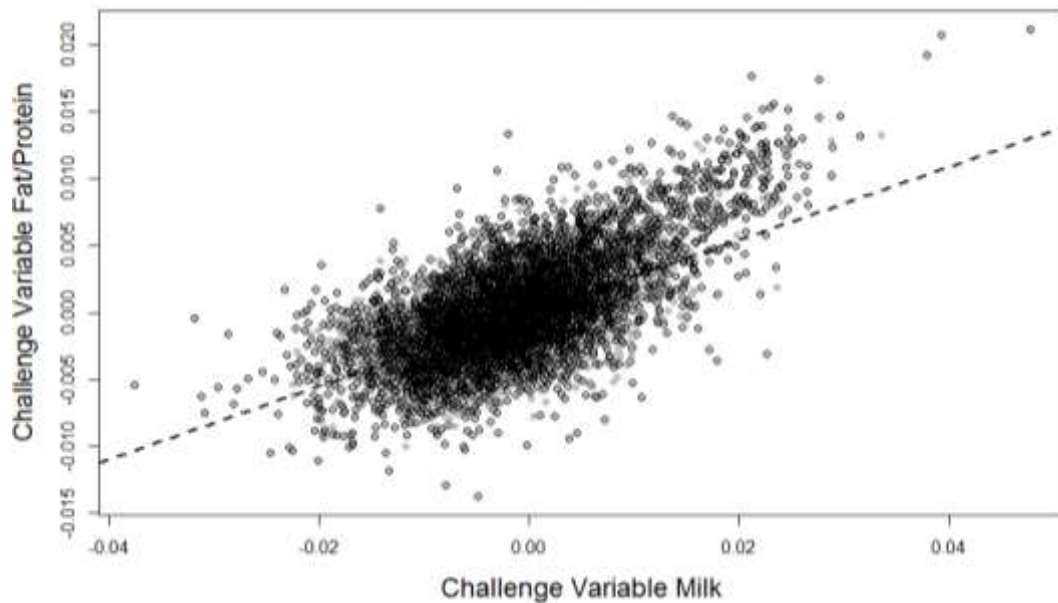


**Figure 4.2:** Empirical distribution of the estimated variance components calculated using the bootstrapping of both random effects and residuals. The red vertical lines depict the estimated values from the original data.

**Table 4.1:** Fixed effects and variance parameter estimates (including the genetic correlation  $r_{sb} = \frac{\sigma_{sb}}{\sqrt{\sigma_s^2 \sigma_b^2}}$ ) from the reaction norm sire model and their estimated standard errors in brackets as obtained from ASReml 3.0 and from the Bootstrap method.

	Parameter Estimates						
	$\mu$	$\beta$	$\sigma_s^2$	$\sigma_b^2$	$\sigma_{sb}$	$r_{sb}$	$\sigma_e^2$
Milk	1.377	-0.0023	0.00423	0.000245	0.000557	0.5478	0.0579
	(0.0080/	(0.0026/	( $1.6 \cdot 10^{-4}$ /	( $2.9 \cdot 10^{-5}$ /	( $5.7 \cdot 10^{-5}$ /	(0.046/	( $1.1 \cdot 10^{-4}$ /
	0.0012)	0.0006)	$1.8 \cdot 10^{-4}$ )	$2.6 \cdot 10^{-5}$ )	$5.2 \cdot 10^{-5}$ )	0.043)	$1.3 \cdot 10^{-4}$ )
Fat/ Protein	1.387	0.0089	0.00435	0.0000515	0.000289	0.6105	0.0579
	(0.0078/	(0.0014/	( $1.7 \cdot 10^{-4}$ /	( $1.1 \cdot 10^{-5}$ /	( $4.1 \cdot 10^{-5}$ /	(0.078/	( $1.1 \cdot 10^{-4}$ /
	0.0012)	0.0004)	$1.8 \cdot 10^{-4}$ )	$9.4 \cdot 10^{-6}$ )	$3.2 \cdot 10^{-5}$ )	0.066)	$1.4 \cdot 10^{-4}$ )

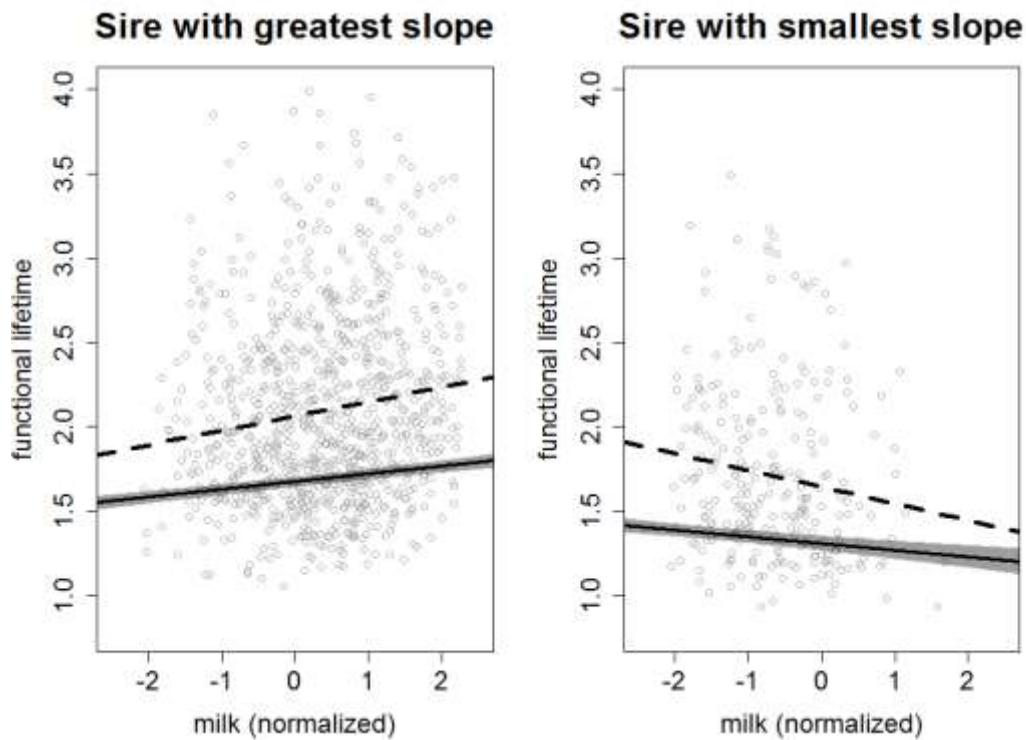
As expected, the analysis of both challenge variables show similar results, which are reflected by the high correlation of 0.608 of the predicted slopes  $b_i$  (compare Figure 4.3) of all bulls. In regard to the reference bulls, the empirical correlation amounts to 0.676, which is significantly different from zero ( $p < 10^{-16}$ ). However, as compared to fat/protein ratio, the analysis of milk yield as challenge variable exhibits greater variation of the slopes. Figure 4.4 displays the fitted regression lines for the two bulls with the greatest and smallest slopes including the 95%-confidence bands obtained from the bootstrap procedure with fixed variance components. We further validated the trends of the daughter performances of the two bulls by means of a simple linear regression of functional lifetime on the challenge variable for each bull, separately (Figure 4.5, dashed lines). As a result, we found significant slope effects of the challenge variable on functional lifetime with p-values of  $2.44 \times 10^{-16}$  and 0.018 for the bulls with the greatest and smallest slopes, respectively.



**Figure 4.3:** Estimated slopes  $b_i$  for the 30,842 animals including the 4,983 reference sires (black circles) for the two challenge variables milk yield and fat to protein ratio.

We further assessed the relationship between the predicted slopes and functional lifetime as well as the two challenge variables. In Figure 4.5, the predicted slope for each sire is

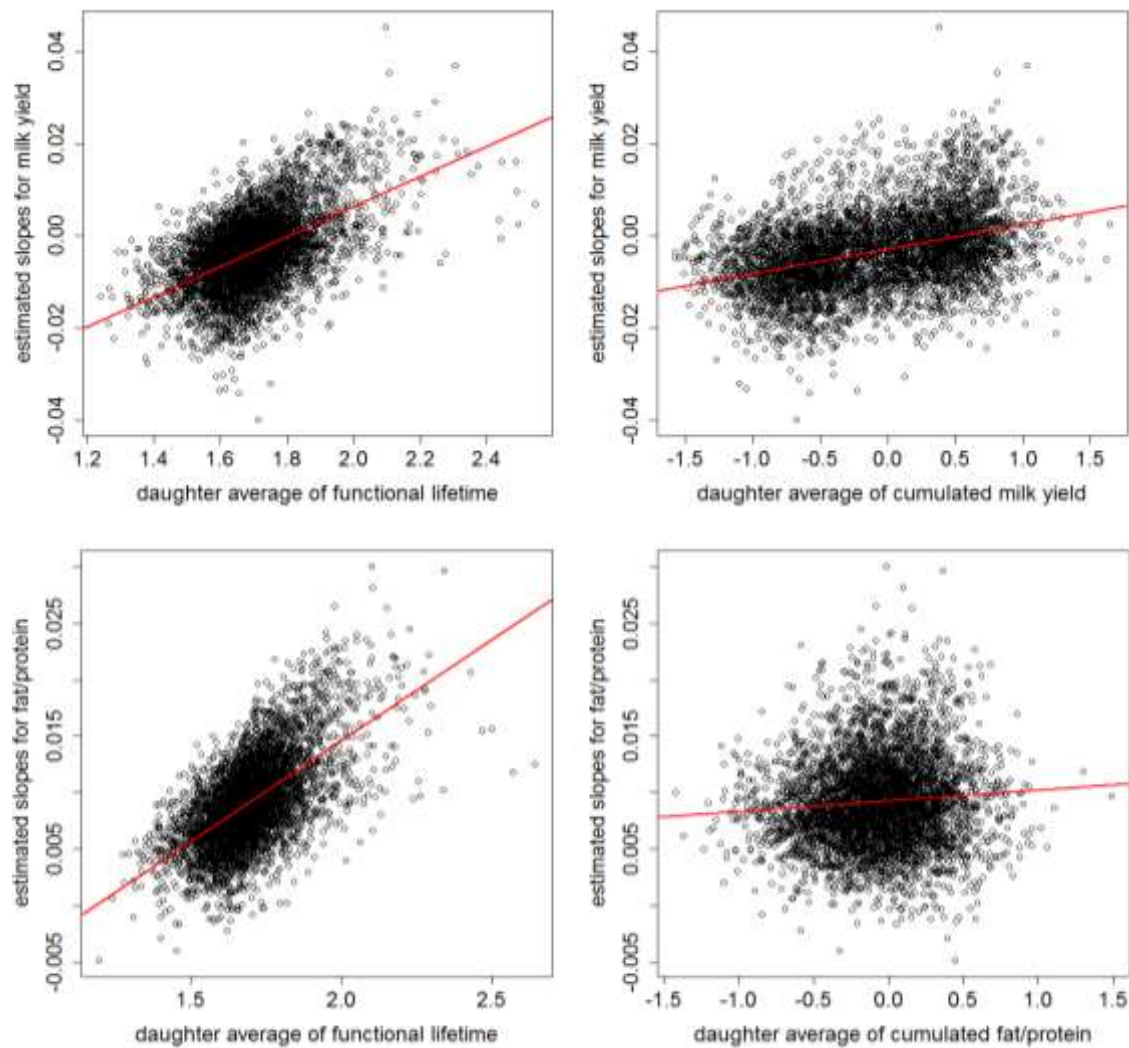
plotted against the average functional lifetime (left panels) or the average challenge (right panels) of all his daughters. The correlation between the slopes and average functional lifetime is 0.535 and 0.61, whereas the correlation between the slopes and averaged challenge amounts to 0.377 and 0.08 for milk yield and fat to protein ratio, respectively.



**Figure 4.4:** The estimated regression lines  $y = \hat{\mu} + \hat{\beta}x + \hat{\sigma}_i + \hat{b}_i x$  for the sire  $i$  with the largest (left) and smallest (right) slopes estimated for the challenge variable milk yield and their corresponding 95% confidence bands (gray) obtained from the bootstrap approach with fixed variance components. The dashed lines depict the corresponding regression lines based on a simple linear regression model fitted for each sire, separately, using the data points of his daughters (points).

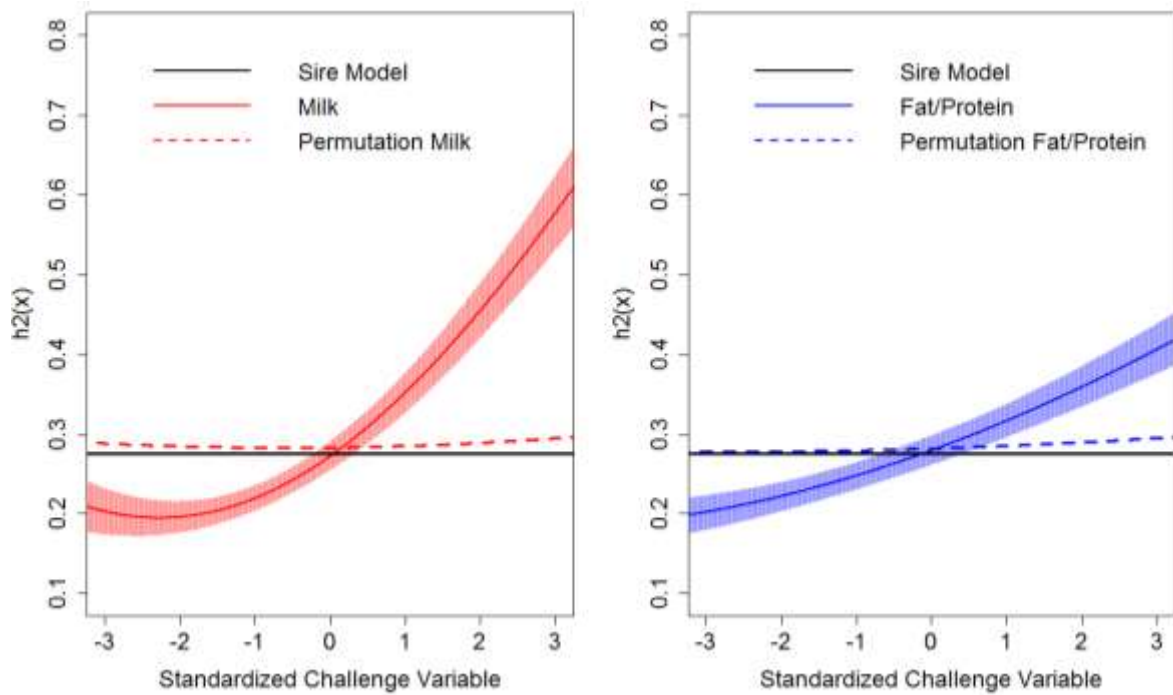
Figure 4.6 illustrates the estimated heritability estimates as functions of the challenge variables. Under the average metabolic load reflected by a challenge of zero, all models including the sire model (2) without any genetic components exhibit similar heritabilities. For both challenge variables, heritability is strongly affected by the challenge. Especially for milk yield, the heritability increases up to ~0.5 for dairy cows that are exposed to extreme

metabolic loads ( $>2$ ). For comparison purposes, we re-estimated the heritability using a reaction norm model based on permuted milk yield as well as permuted fat to protein ratio data as challenge variable. As expected, the influence of the challenge load on the heritability is substantially diminished.



**Figure 4.5:** The estimated slopes for the challenge variable milk yield in dependency of the daughter average of functional lifetime (upper left) and the cumulated milk yield (upper right) and the corresponding plots for fat/protein ratio (below left and right). The red line depicts the linear regression of the y-axis to the x-axis.





**Figure 4.6:** Estimated heritabilities of functional lifetime as functions of the challenge variables and their corresponding 95% confidence bands as obtained from the non-parametric bootstrap method. The two dashed lines represent the heritabilities estimated from a reaction norm model based on permuted milk yield or fat to protein ratio data.

## Discussion

The objective of the present study was the introduction and evaluation of a new measure for the metabolic robustness of dairy cows. Our hypothesis is that some cows are genetically better suited to cope with the metabolic load during early lactation than other cows regardless of the extent of the metabolic load. We further assumed that the metabolic robustness of these cows is reflected by a greater fitness, e.g. lower frequency of disorder incidences, which in the long run will result in a longer functional lifetime. To this end, we proposed a reaction norm sire model linking the functional lifetime to the metabolic load in the early part of the first lactation.

The metabolic load or stress that a certain cow encounters during early lactation is best characterized by her energy balance (EB) [32], which is mainly negative during early lacta-

tion. The EB, however, is very difficult and costly to practically measure [33] and is currently not implemented in routine genetic evaluation systems. Therefore, in our study, we made use of the accumulated milk yield or fat to protein ratio of the first three test day records as indicator challenge variables and proxy for the negative EB. Indeed, the fat to protein ratio in the early part of the first lactation has been found to be a reliable indicator for EB [33–35] due to their high phenotypic [36] as well as genetic correlations [33]. Moreover, Heuer et al. [35] discovered that a higher fat to protein ratio is associated with a higher risk for production-related diseases. They also reported that cows with higher milk yields are more prone to develop metabolic disorders [35] and thus milk yield as well as fat to protein ratio are well-suited indicators for the metabolic stress of dairy cows at least during early lactation ( $\text{DIM} < 180$ ) [33].

The results of our study have shown that metabolic robustness of dairy cows has a genetic component. For both challenge variables, a significant proportion of the sire variation for the slope of the random regression is explained by the genetic relatedness of the bulls, suggesting a genetic component for the ability to cope with metabolic stress. In other words, some bulls have daughters in which metabolic stress (as indicated by either milk yield or fat to protein ratio) is associated with a reduction of longevity, while for other bulls the longevity of the daughters remains largely unaffected by an energy deficit in the transition phase.

In the present study, two measures of robustness were calculated based on two different challenge variables that were supposed to reflect the metabolic load. Interestingly, we find a high correlation between the two measures (Figure 4.3) supporting the suitability of the two challenge variables as proxy for the metabolic load. Nevertheless, according to our results, using cumulated milk yield as challenge yielded a higher slope variation and a higher heritability for longevity at least for dairy cows exposed to high challenges (com-

pare Figure 4.6). The substantial variation of the slopes also exceeded our initial expectations, which are illustrated in Figure 4.1. According to our hypothesis, robust dairy cows are expected to be largely unaffected by the metabolic load. However, in our results we observe bulls whose daughters' functional lifetime is actually positively influenced by metabolic stress (Figure 4.4, left panel). A possible explanation for this could be a special treatment of high-yielding cows. If a high-yielding dairy cow became ill, farmers might put more effort into aiding her recovery than in the case of other cows producing less milk.

In general, reaction norm models aim to model genotype by environment interactions given that the environment is measureable on a continuous scale [19]. In cattle, reaction norm models have been successfully applied to describe the reaction of different milk production and fertility traits to changing environments [37–39]. In this setting, the slope is a measure of phenotypic sensitivity to environmental changes and may be viewed as an own trait [17,40]. In our study, the environment was defined to be the metabolic load during early lactation and is thus a kind of 'internal' environment. Besides the metabolism and performance of dairy cows, environmental factors such as management, housing and herd also have a high impact on metabolic stress [4].

Kolmodin and Bijma [17] studied the special features for selection, when the breeding goal is a desired gain in intercept or slope. They pointed out that the optimum selection environment is dependent on several factors including the correlation between slope and intercept and the heritability of the trait. Therefore, special attention needs to be paid to these features when using the suggested slopes for the breeding of metabolically robust dairy cows. Applying their theory to our model and considering the varying heritability in Figure 4.6, we conclude that in order to obtain the highest selection response in the slopes, we have to perform selection among cows exposed to the highest metabolic challenges. Due to

the positive correlation of intercept and slope, this would lead to the selection of cows that are both long-living and high-challenged and thus considered metabolically robust.

The transition phase of dairy cows has long been an unresolved problem for the farmer industry. Increasing health and fertility problems have been observed worldwide as a consequence of the breeding for high-producing dairy cows [4,33]. Therefore, the hypothesis of a genetic component of the metabolic robustness has been of major interest recently, not least because of its potential for the dairy industry to improve the transition cow's health without compromising milk production. To our knowledge, this analysis represents the first study trying to assess the metabolic robustness of dairy cows using reaction-norm models. Even though the model was only fit to data of the Brown Swiss population, we believe that our approach is also applicable to other breeds leading to similar results.

## References

1. Gross J, van Dorland HA, Bruckmaier RM, Schwarz FJ. Performance and metabolic profile of dairy cows during a lactational and deliberately induced negative energy balance with subsequent realimentation. *J. Dairy Sci.* 2011;94:1820–30.
2. Bobe G, Young JW, Beitz DC. Invited review: pathology, etiology, prevention, and treatment of fatty liver in dairy cows. *J. Dairy Sci.* 2004;87:3105–24.
3. Gross JJ, Schwarz FJ, Eder K, van Dorland HA, Bruckmaier RM. Liver fat content and lipid metabolism in dairy cows during early lactation and during a mid-lactation feed restriction. *J. Dairy Sci.* 2013;96:5008–17.
4. Sundrum A. Metabolic Disorders in the Transition Period Indicate that the Dairy Cows' Ability to Adapt is Overstressed. *Animals.* 2015;5:978–1020.
5. Greenfield RB, Cecava MJ, Donkin SS. Changes in mRNA expression for gluconeogenic enzymes in liver of dairy cattle during the transition to lactation. *J. Dairy Sci.* 2000;83:1228–36.
6. Drackley JK, Overton TR, Douglas GN. Adaptations of Glucose and Long-Chain Fatty Acid Metabolism in Liver of Dairy Cows during the Periparturient Period. *J. Dairy Sci.* 2001;84, Supplement:E100–12.

7. Reynolds CK, Aikman PC, Lupoli B, Humphries DJ, Beever DE. Splanchnic metabolism of dairy cows during the transition from late gestation through early lactation. *J. Dairy Sci.* 2003;86:1201–17.
8. Loor JJ, Dann HM, Everts RE, Oliveira R, Green CA, Guretzky NAJ, et al. Temporal gene expression profiling of liver from periparturient dairy cows reveals complex adaptive mechanisms in hepatic function. *Physiol. Genomics.* 2005;23:217–26.
9. Drackley JK, Dann HM, Douglas GN, Guretzky NAJ, Litherland NB, Underwood JP, et al. Physiological and pathological adaptations in dairy cows that may increase susceptibility to periparturient diseases and disorders. *Growth.* 2005;7.
10. Graber M, Kohler S, Kaufmann T, Doherr MG, Bruckmaier RM, van Dorland HA. A field study on characteristics and diversity of gene expression in the liver of dairy cows during the transition period. *J. Dairy Sci.* 2010;93:5200–15.
11. Schlegel G, Ringseis R, Keller J, Schwarz FJ, Eder K. Changes in the expression of hepatic genes involved in cholesterol homeostasis in dairy cows in the transition period and at different stages of lactation. *J. Dairy Sci.* 2012;95:3826–36.
12. Ostrowska M, Górka BŻ, Śloniewski K, Kowalski LZ. Expression of PC, PCK1, PCK2, LDHB, FBP1 and G6PC genes in the liver of cows in the transition from pregnancy to lactation. *Anim. Sci. Pap. Rep.* 2013;31:281–90.
13. Ha N-T, Gross JJ, van Dorland A, Tetens J, Thaller G, Schlather M, et al. Gene-based mapping and pathway analysis of metabolic traits in dairy cows. *PloS One.* 2015;10:e0122325.
14. Kessel S, Stroehl M, Meyer HHD, Hiss S, Sauerwein H, Schwarz FJ, et al. Individual variability in physiological adaptation to metabolic stress during early lactation in dairy cows kept under equal conditions. *J. Anim. Sci.* 2008;86:2903–12.
15. Gross JJ, R M Bruckmaier. Repeatability of metabolic responses to a nutrient deficiency in early and mid-lactation and its implication to robustness of dairy cows. *J. Dairy Sci.* 2015;
16. Ingvarlsen KL, Dewhurst RJ, Friggens NC. On the relationship between lactational performance and health: is it yield or metabolic imbalance that cause production diseases in dairy cattle? A position paper. *Livest. Prod. Sci.* 2003;83:277–308.
17. Kolmodin R, Bijma P. Response to mass selection when the genotype by environment interaction is modelled as a linear reaction norm. *Genet. Sel. Evol. GSE.* 2004;36:435–54.
18. Streit M, Reinhardt F, Thaller G, Bennewitz J. Reaction norms and genotype-by-environment interaction in the German Holstein dairy cattle. *J. Anim. Breed. Genet.* 2012;129:380–9.

19. Streit M, Reinhardt F, Thaller G, Bennewitz J. Genome-wide association analysis to identify genotype× environment interaction for milk protein yield and level of somatic cell score as environmental descriptors in German Holsteins. *J. Dairy Sci.* 2013;96:7318–24.
20. Mészáros G, Sölkner J, Ducrocq V. The Survival Kit: Software to analyze survival data including possibly correlated random effects. *Comput. Methods Programs Biomed.* 2013;110:503–10.
21. Joaquim T, Jesús P, Vincent D. Validation of an approximate approach to compute genetic correlations between longevity and linear traits. *Genet Sel Evol.* 2006;38:65–83.
22. Ducrocq V, Delaunay I, Boichard D, Mattalia S. A general approach for international genetic evaluations robust to inconsistencies of genetic trends in national evaluations. *Interbull Bull.* 2003;101.
23. Gilmour AR, Gogel BJ, Cullis BR, Thompson R, Butler D. ASReml user guide release 3.0. 2009;
24. Strandberg E, Kolmodin R, Madsen P, Jensen J, Jorjani H. Genotype by environment interaction in Nordic dairy cattle studied by use of reaction norms. *Interbull Bull.* 2000;41.
25. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. *Biometrics.* 1994;1171–7.
26. Thai H-T, Mentré F, Holford NHG, Veyrat-Follet C, Comets E. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharm. Stat.* 2013;12:129–40.
27. Van der Leeden R, Busing F, Meijer E. Bootstrap methods for two-level models. *Multilevel Conf.* 1997.
28. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge university press; 1997.
29. Carpenter JR, Goldstein H, Rasbash J. A novel bootstrap procedure for assessing the relationship between class size and achievement. *J. R. Stat. Soc. Ser. C Appl. Stat.* 2003;52:431–43.
30. MacKinnon JG. Bootstrap methods in econometrics\*. *Econ. Rec.* 2006;82:S2–18.
31. Wang L. A genetic variation of the p38 $\beta$  promoter region is correlated with an increased risk of sporadic colorectal cancer. *Oncol. Lett.* [Internet]. 2013 [cited 2013 Aug 16]; Available from: <http://www.spandidos-publications.com/10.3892/ol.2013.1334>

32. Coffey MP, Emmans GC, Brotherstone S. Genetic evaluation of dairy bulls for energy balance traits using random regression. *Anim. Sci.-Glasg.-*. 2001;73:29–40.
33. Buttchereit N, Stamer E, Junge W, Thaller G. Short communication: Genetic relationships among daily energy balance, feed intake, body condition score, and fat to protein ratio of milk in dairy cows. *J. Dairy Sci.* 2011;94:1586–91.
34. Grieve DG, Korver S, Rijpkema YS, Hof G. Relationship between milk composition and some nutritional parameters in early lactation. *Livest. Prod. Sci.* 1986;14:239–54.
35. Heuer C, Schukken YH, Dobbelaar P. Postpartum Body Condition Score and Results from the First Test Day Milk as Predictors of Disease, Fertility, Yield, and Culling in Commercial Dairy Herds. *J. Dairy Sci.* 1999;82:295–304.
36. Buttchereit N, Stamer E, Junge W, Thaller G. Evaluation of five lactation curve models fitted for fat: protein ratio of milk and daily energy balance. *J. Dairy Sci.* 2010;93:1702–12.
37. Kolmodin R, Strandberg E, Madsen P, Jensen J, Jorjani H. Genotype by Environment Interaction in Nordic Dairy Cattle Studied Using Reaction Norms. *Acta Agric. Scand. Sect. — Anim. Sci.* 2002;52:11–24.
38. Calus MPL, Groen AF, de Jong G. Genotype  $\times$  Environment Interaction for Protein Yield in Dutch Dairy Cattle as Quantified by Different Models. *J. Dairy Sci.* 2002;85:3115–23.
39. Hayes BJ, Carrick M, Bowman P, Goddard ME. Genotype  $\times$  environment interaction for milk production of daughters of Australian dairy sires from test-day records. *J. Dairy Sci.* 2003;86:3736–44.
40. de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 2002;9:67–103.

## **CHAPTER 5 GENERAL DISCUSSION**



## Overview

In order to assess the hypothesis of a genetic component of the metabolic adaptation, the present dissertation comprises three independent studies aiming to analyse this hypothesis from different point of views. In **chapter 2**, we addressed the metabolic adaptation by means of auxiliary phenotypes known to be important for the process of adaptation. We identified a number of significant genes and pathways affecting some of the auxiliary traits, which may be viewed as potential candidates for the genetic basis of the metabolic adaptation. Another pool of candidate genes and pathways was derived from a transcriptomic perspective in **chapter 3**. Here, we used RNA-seq data to identify differentially expressed genes and pathways before and after lactation. The basis of the genomic and transcriptomic analyses in **chapter 2** and **chapter 3**, respectively, is the metabolic adaptation, which is a complex process rather than a measureable phenotype. If a genetic component exists, which affects this process of adaptation, this should be reflected in a better adaptability of certain dairy cows that are considered metabolically robust. In order to measure this robustness as a phenotype, in **chapter 4** we proposed a reaction norm sire model to predict the metabolic robustness of Brown Swiss bulls based on the metabolic performance of their daughters. In all three studies, we found substantial evidence supporting the genetic basis of the metabolic adaptation by means of different methods and independent data sets. Even though genome-wide association studies (GWAS), differential gene expression analysis (DGE analysis) and reaction norm models are well established methods, there are still some limitations and challenges in their applications, which we will discuss in the following.

## Genome Annotations – the Bridge Between Genome and Functional Biology?

A crucial step in the gene-based GWAS and DGE analysis in **chapter 2** and **3** is the annotation of genes. In **chapter 2**, SNPs were grouped into sets according to their bp positions in relation to annotated genes, whereas in **chapter 3**, mapped reads were counted if they fell into annotated gene regions. In both studies, the gene annotation was retrieved from the Ensembl database [1]. The Ensembl project is a joint project between the European Bioinformatics Institute (EMBI-EBI) and the Wellcome Trust Sanger Institute (WTSI). The major goal of gene annotation is to bridge the gap between the genome and the proteome, i.e. the totality of all proteins expressed in a species. In practice, this is done by aligning known protein sequences to a reference genome, both originating from publicly available resources. This annotation along with the functional information on specific gene products, i.e. proteins, is the key element for a successful interpretation of our results and was treated as reliable information.

However, despite the substantial progress in technology as well as improvement of bioinformatic tools, the recent available genome assemblies as well as protein databases are far from completed at least for the species *Bos taurus*. For instance, in the two most frequently used and currently available genome assemblies (UMD3.1.1 [2] and Btau\_4.6 [3]) ~10% of the sequences are missing with numerous gaps that still have to be filled. Another source of uncertainty may also arise from the protein sequences. Based on a query in the UniProt Knowledgebase (UniProtKB, [www.uniprot.org/uniprot/](http://www.uniprot.org/uniprot/)), one of the protein databases employed by Ensembl, we found a total of ~24,000 records of proteins for the species *Bos taurus*, of which only a fraction of ~6,000 proteins are validated and reviewed. Hence, the quality of the resulting gene annotation is dependent of three main points: (1) the reliability of the reference genome, (2) the reliability of the protein sequences and last but not least

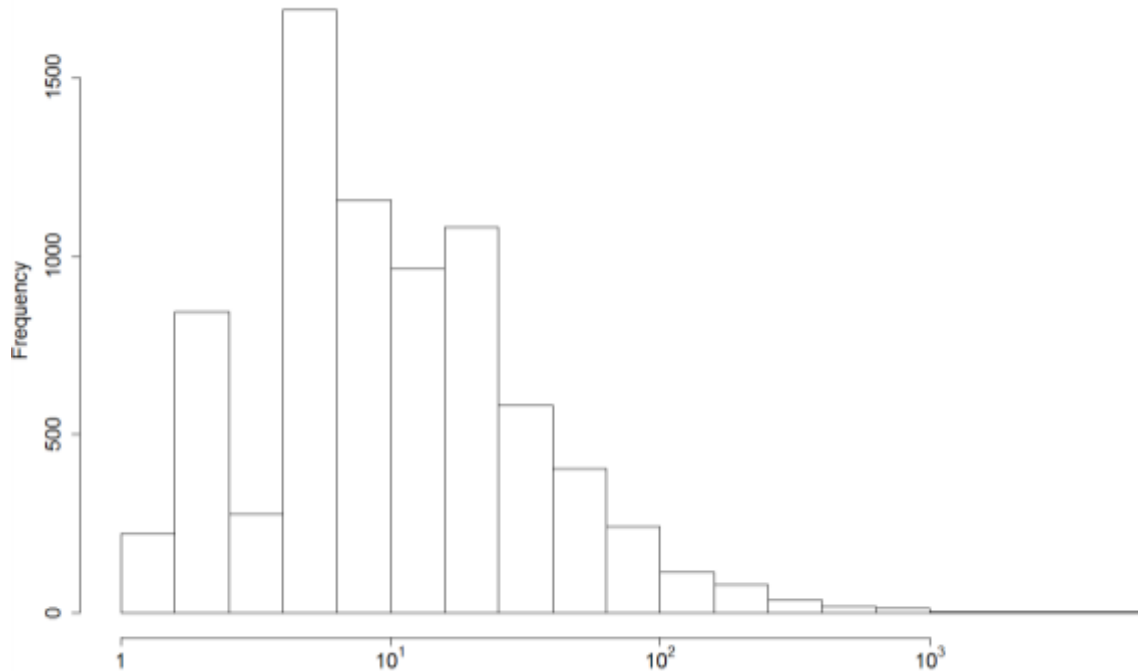
(3) the procedure of mapping these two sources of information. Each of the three components is important for the reliability of the gene annotation and thus highly influenced the precision of biological analyses in **chapter 2** and **chapter 3**.

How these uncertainties may relate to our results may be predicted from the experiences made in human data. Within a time frame of three years the first human assembly containing ~150,000 gaps with 10% of the genome missing was improved to the ‘nearly’ complete assembly with only ~350 gaps and 1% missing [2,4]. With this improvement, they detected that 58% of the transcripts mapped onto the ‘old’ genome assembly were erroneous [4]. Indeed, in **chapter 3**, about ~30% of our RNA-seq data (read pairs) were not included in the final analysis, since we were either not able to map it to the reference genome or assign them to any gene products. This limitation, however, is common to the majority of RNA-seq analyses for the cattle genome, for example in McCabe et al. [5] only a fraction of ~30% were retained for final analyses after several steps of quality controls, whereas in McLoughlin et al. [6] ~50% of the mapped reads fell into annotated gene regions. This together with the results of our DGE analysis in **chapter 3** demonstrated the incompleteness of the current genome annotation. Therefore, future studies may aim at unveiling this ‘missing’ part of the bovine transcriptome either by de-novo alignment of the transcriptome or on ‘updated’ genome annotations.

### **RNA-seq – the Curse of Small Sample Sizes**

Another shortcoming of the RNA-seq experiment in **chapter 3** is the extremely low number of samples. As a matter of fact, next-generation sequencing technology is still a relatively expensive tool. Consequently, small sample sizes are common practice in RNA-seq experiments, as demonstrated in Figure 5.1. The number of samples illustrated in this figure refers to the total number samples in a given experiment with a median of nine samples per experiment. Typically, RNA-seq experiments are suspected to compare samples among

two or more conditions and thus, the effective number of samples per condition is even smaller. Therefore, comparison studies have been conducted recently to assess the general effect of small sample sizes in RNA-seq studies [7,8]. On the one hand, Seyednasrollah et al. [7] observed an appropriate power when the number of replicates, i.e. the number of samples per condition, is greater than or equal to five. On the other hand, Schurch et al. [8] pointed out that for the detection of highly differentially expressed genes (e.g. fold change  $> 2$ ), three replicates may be sufficient, whereas an appropriate power to detect ~90% of true effects would only be possible with more than 12 replicates. Nevertheless, as demonstrated in **chapter 3**, many of our results were matched by previous studies and could be plausibly related to the transition dairy cow. However, considering the results of Schurch et al. [8], we might still have missed some further important genes that may also have an important influence of the hepatic adaptation of dairy cows.



**Figure 5.1:** Histogram for the number of samples in RNA-seq experiments ( $\log_{10}$ ). These numbers were obtained by a query from the NCBI database (<http://www.ncbi.nlm.nih.gov/gds>) for all GEO datasets of type "expression profiling by high throughput sequencing" (date of query: 5.5.2016).

## Reaction Norm Model – the Model Assumptions

Compared to **chapter 3**, in **chapter 4** a dataset of more than 500,000 records was used to fit a reaction norm sire model and to predict a measure of metabolic robustness of around 5,000 bulls. Despite the high amount of data available, major challenges of this study lay in the assumption of a linear relationship between response and challenge variables and the choice of these variables. As stated in **chapter 1**, reaction norm models usually assume a linear relationship between response and challenge variables (e.g. [9–12]), not least because the interpretability of the slope as the environmental sensitivity or in our case, as the metabolic robustness that we are seeking for. Nevertheless, nonlinear relationships between functional longevity and other traits exist in dairy cattle, for instance conformation traits [13], type traits or somatic cell score [14]. In order to test whether a nonlinear relationship also exists for longevity and the production traits in **chapter 4**, we used the Akaike Information Criterion [15] to assess whether inclusion of higher order polynomials would improve our statistical model and found indications for higher order relationships (data not shown). Taking this effect into consideration could further increase the estimated genetic variances. We therefore believe that the estimated genetic merit of the metabolic robustness in **chapter 4** might still be underestimated.

With respect to our definition of the metabolic robustness as the general impact of the metabolic load during early lactation on the cow's fitness, longevity is the most appropriate response variable to choose, since it includes the whole fitness complex [16]. However, in practice, measurements of the true longevity of a cow, generally defined as the productive period following first parity until disposal, are extremely biased due to the voluntary culling of cows exhibiting poor production [16]. This has posed a problem in many genetic evaluations in the past (e.g. [17,18]). Therefore, it is a common practice to consider the 'functional' instead of the true longevity, in which the effect of voluntary culling on milk

production and other environmental factors should theoretically be adjusted for [17,19]. However, as demonstrated by Dekkers [18], in practice this kind of adjustment is still imperfect. This observation is also in agreement with our observation in **chapter 4**, where a positive relationship between milk yield and functional lifetime was observed for some bulls. Hence, genetic analyses involving longevity data should always be interpreted with caution.

In **chapter 4**, cumulated milk yield and fat to protein ratio in the early part of first lactation were used as indicators for the extent of the metabolic load. A robustness score is then predicted based on the relationship between the longevity and the metabolic load in the first lactation, ignoring the possibility that follow-up lactations may also have an impact on the cow's functional lifetime. An alternative would be to consider the average metabolic load (e.g. cumulated milk yield (first three records) averaged over all parities); however, this might lead to a bias in the number of parities, which differs substantially between cows. Instead of changing the challenge variable, another possibility would be to redefine the response, e.g. by considering the probability of surviving the first lactation as a direct effect of its metabolic load or the frequency of disorder occurrences.

## Outlook

Despite the mentioned limitations of the methods applied in our three studies, we found substantial evidence supporting the hypothesis of a genetic basis of the metabolic adaptation. Following the suggestions of further studies on this genetic basis, for example when a more comprehensive genome annotation is available or with better defined response variables, we believe that the disclosure of this genetic basis will be possible. As a final goal, we would like to find an overlap between the current three studies by means of a GWAS for the newly defined metabolic robustness in **chapter 4** to determine whether the identi-

fied genetic factors are included in the pools of candidate genes obtained in **chapter 2** and **chapter 3**.

## References

1. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. *Nucleic Acids Res.* 2016;44:D710–6.
2. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 2009;10:R42.
3. Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science.* 2009;324:522–8.
4. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. *Nature.* 2004;431:931–45.
5. McCabe M, Waters S, Morris D, Kenny D, Lynn D, Creevey C. RNA-seq analysis of differential gene expression in liver from lactating dairy cows divergent in negative energy balance. *BMC Genomics.* 2012;13:193.
6. McLoughlin KE, Nalpas NC, Rue-Albrecht K, Browne JA, Magee DA, Killick KE, et al. RNA-seq transcriptional profiling of peripheral blood leukocytes from cattle infected with *Mycobacterium bovis*. *Dev. Bov. Immunol.- Integr. View.* 2015;10:7.
7. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* 2015;16:59–70.
8. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA [Internet].* 2016 [cited 2016 May 5]; Available from: <http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.053959.115>
9. Streit M, Reinhardt F, Thaller G, Bennewitz J. Genome-wide association analysis to identify genotype× environment interaction for milk protein yield and level of somatic cell score as environmental descriptors in German Holsteins. *J. Dairy Sci.* 2013;96:7318–24.
10. Hayes BJ, Carrick M, Bowman P, Goddard ME. Genotype x environment interaction for milk production of daughters of Australian dairy sires from test-day records. *J. Dairy Sci.* 2003;86:3736–44.
11. Calus MPL, Groen AF, de Jong G. Genotype × Environment Interaction for Protein Yield in Dutch Dairy Cattle as Quantified by Different Models. *J. Dairy Sci.* 2002;85:3115–23.

12. Kolmodin R, Strandberg E, Madsen P, Jensen J, Jorjani H. Genotype by Environment Interaction in Nordic Dairy Cattle Studied Using Reaction Norms. *Acta Agric. Scand. Sect. — Anim. Sci.* 2002;52:11–24.
13. Zavadilová L, Němcová E, Štípková M, Bouška J. Relationships between longevity and conformation traits in Czech Fleckvieh cows. *Czech J Anim Sci.* 2009;54:387–94.
14. Sewalem A, Miglior F, Kistemaker GJ, Van Doormaal BJ. Analysis of the Relationship Between Somatic Cell Score and Functional Longevity in Canadian Dairy Cattle. *J. Dairy Sci.* 2006;89:3609–14.
15. Akaike H. A new look at the statistical model identification. *IEEE Trans. Autom. Control.* 1974;19:716–23.
16. Essl A. Longevity in dairy cattle breeding: a review. *Livest. Prod. Sci.* 1998;57:79–89.
17. Essl A. Estimation of the genetic correlation between first lactation milk yield and length of productive life by means of a half-sib analysis: A note on the estimation bias. *J. Anim. Breed. Genet.* 1989;106:402–8.
18. Dekkers JC. Theoretical basis for genetic parameters of herd life and effects on response to selection. *J. Dairy Sci.* 1993;76:1433–43.
19. Joaquim T, Jesús P, Vincent D. Validation of an approximate approach to compute genetic correlations between longevity and linear traits. *Genet Sel Evol.* 2006;38:65–83.



## Acknowledgements

My thanks go to Prof. Henner Simianer and Prof. Martin Schlather for being my supervisors and for their great support during my whole PhD study. I would especially like to express my gratitude to Prof. Henner Simianer, for his continuing guidance, valuable advice and inspiring ideas during all this time and for his great patience and encouragement whenever I felt down. All this would have been not possible, if it were not for Prof. Martin Schlather, who introduced me to the great working group of Prof. Henner Simianer and thus, I would like to also thank you, Martin, for this opportunity and also for your valuable advices whenever I needed it.

I would like to thank Prof. Rupert Bruckmaier, Ms Aebi, Fritz Schmitz-Hsu and Josef Groß for the great collaboration and their precious support during the whole project. In particular, I want to thank you, Josef, for always being available for my questions and concerns about physiological aspects and for always welcoming me in Bern, giving me the opportunity to visit such a wonderful city.

I also would like to thank Prof. Thomas Kneib for being my third supervisor and for giving me the great opportunity to be an associate member of the research training group ‘Scaling Problems in Statistics’ (RTG 1644) supported by the Deutsche Forschungsgemeinschaft (DFG).

I also thank the Swiss Commission for Technology and Innovation CTI (project no. 13948.2 PFLS-LS) and Swissgenetics (Zollikofen, Switzerland) for their grants, which made this research project possible.

I am also grateful to all my colleagues from the Animal Breeding and Genetics Group for their company and the ‘regular coffee breaks’ making my working environment much more fruitful and fun. Especially to you, Ms Döring, Christian, Anna, Reza, Svetlana, Mahmood and Michael, thank you for bearing my temperament, for cheering me up whenever I felt down and for your friendship.

Last but not least, I would like to thank my whole family: Thank you, mum and dad, for your unconditional love and support, which gave me the strength to pull through to the end. Thank you, my lovely sisters and brothers, for always encouraging me, sharing joys and sorrows and finally, thank you, ‘Anh’ and my little boy Dan, for always being on my side and sharing your whole lives with me.

## Curriculum vitae

Name: Ngoc-Thuy Ha  
 Date of birth: January 12, 1988  
 Place of birth: Ben-Tre / Vietnam  
 Nationality: German

## Education

Since 02/2013 **Ph.D Student in Agricultural Sciences at Georg-August University, Göttingen**  
 Thesis title: “Lost in transition - Genetic, Transcriptomic and Breeding Aspects of Metabolic Robustness in Dairy Cows”  
 Thesis advisor: Prof. Dr. Henner Simianer

10/2010 – 11/2012 **Master of Science with Distinction in Mathematics at Georg-August University, Göttingen**  
 Final mark: 1.3  
 Major field of study: Mathematical Stochastics, final mark: 1.0  
 Master’s thesis: “Tail Estimation Methods for the Number of False Positive Associations in the Drosophila Genetic Reference Panel”, mark: 1.0

10/2007 – 09/2010 **Bachelor of Science in Mathematics at Georg-August University, Göttingen**  
 Final mark: 1.8  
 Major field of study: Mathematical Stochastics, final mark: 1.0  
 Bachelor’s thesis: “Bootstrap, Maximum-Likelihood und Bayes-Methoden für Regressionsmodelle” (English: “Bootstrap, Maximum-Likelihood and Bayes-Methods for Regression Models”), mark: 1.0

## Professional Experience

02/2013 – 04/2016 **Scientific Assistant at the Veterinary Physiology, Prof. Dr. Rupert Bruckmaier, Vetsuisse Faculty of University Bern**  
 Researched on the metabolic robustness of dairy cows (project “Only Robust”) in collaboration with the Georg-August University Göttingen, Prof. Dr. Henner Simianer

10/2010 – 11/2012 **Student Assistant at the Department of Genetic Epidemiology, Professor Bickeböller, Georg-August University, Göttingen**

Researched on an own project “Coverage on current SNP Chips”  
and assisted PhD students with their research

04/2009 – 09/2010     **Student Assistant at the Centre for Statistics, Georg-August University, Göttingen**

### **Voluntary Activities**

09/2015     **Attendance at the 66<sup>th</sup> Annual Meeting of the European Federation of Animal Science, Warsaw**

Talk “Gene and pathway analysis of metabolic traits in dairy cows”

09/2015     **Attendance at the Deutsche Gesellschaft für Züchtungskunde e. V. (DGfZ) Meeting, Berlin**

Talk “Genetische Analyse der Stoffwechselstabilität von frischlaktierenden Milchkühen”

04/2012     **Attendance at the European Mathematical Genetics Meeting 2012, Göttingen**

Poster “Gene and Local Coverage on Current SNP chips”

06/2012     **Attendance at the Research Seminar “New Developments in Quantitative and Population Genetics of Drosophila”, Göttingen**

Talk “Tail Estimation Methods

for the Number of False Positive Associations in the Drosophila Genetic Reference Panel”

### **Peer-reviewed Papers**

**Ha, Ngoc-Thuy**, Saskia Freytag, und Heike Bickeboeller. „Coverage and Efficiency in Current SNP Chips“. *European Journal of Human Genetics*, 22. Januar 2014. doi:10.1038/ejhg.2013.304.

**Ha, Ngoc-Thuy**, Josef Johann Gross, Annette van Dorland, Jens Tetens, Georg Thaller, Martin Schlather, Rupert Bruckmaier, und Henner Simianer. „Gene-Based Mapping and Pathway Analysis of Metabolic Traits in Dairy Cows“. *PLoS ONE* 10, Nr. 3 (19. März 2015): e0122325. doi:10.1371/journal.pone.0122325.